Variational Approximations and How to Improve Them

Simon Barthelme, Gipsa-lab, CNRS

23rd October 2018

Outline

- Brief reminder on variational methods
- What's worth improving?
- How to improve: importance sampling
- How to improve: perturbations
- Open questions

Disclaimer

Very little of this is novel or my own work, but (therefore?) I think it's useful. Main ideas appear in Tierney, Kass & Kadane (1989). Lots of similar ideas appeared earlier in the stat. physics literature, see Opper & Winther (2001).

A reminder

- The goal of variational methods in Bayesian statistics is to compute an approximation to the posterior distribution (or just to posterior moments).
- Should be cheaper than sampling.
- Sometimes variational methods are incredibly accurate, sometimes the results are really bad.
- There are dozens of methods out there, I'm just going to talk about general principles that apply whenever you have an initial approximation and you wish to improve it.

Goal of this tutorial

General setup: you've already computed an approximation and:

- You want to squeeze a bit more accuracy out of the method
- You want to extract some higher-moments or marginals
- (You wish to know how good the approximation is)

Setup, notation

Let $\pi(\theta) \propto p(\theta|y)$ (posterior distribution over parameter θ given data y). We call π the "target distribution". $q(\theta)$ is an approximation to π that belongs to a family of approximating distributions $\mathcal{Q}$. Typically $\mathcal{Q}$ is the set of Gaussian distributions, so we are looking for a Gaussian approximation to the posterior. I'll focus on the univariate case just to simplify notation.

Variational Bayes in one slide

In VB we find an approximation that minimises the following cost:

$$\operatorname*{argmin}_{q \in \mathcal{Q}} KL(q||\pi)$$

This is tractable because $KL(q||\pi)$ is an expectation over q. To obtain the various flavours:

- Change the approximating family $\mathcal{Q}$
- Change how you do the optimisation
- Change the cost function (rinse, repeat)

In some cases VB=mean field approximation from statistical physics.

Expectation Propagation

Expectation Propagation (Minka, 2001) is another algorithm that works on different principles, essentially by refining approximations to each likelihood term in the posterior:

$$\pi(\theta) = \prod_{i=1}^{n} p(y_i|\theta)p(\theta) \approx q(\theta) = \prod_{i=1}^{n} q_i(\theta)p(\theta)$$

The well-known "belief propagation" algorithm is a special case of EP.

The Canonical Gaussian Approximation

The CGA is sometimes (mistakenly) called the "Laplace" approximation. In two steps:

1. Find the mode of the target

$$\theta^\star = \underset{\theta}{\text{argmax}} \log \pi(\theta)$$

2. Compute second derivative of $\log \pi$ at the mode

$$h = -\frac{\partial^2}{\partial \theta^2} \log \pi(\theta)|_{\theta=\theta^\star}$$

The CGA is the Gaussian distribution centred at θ with variance h^{-1}.

Running example

Running example for this talk: a sequence of Bernoulli trials

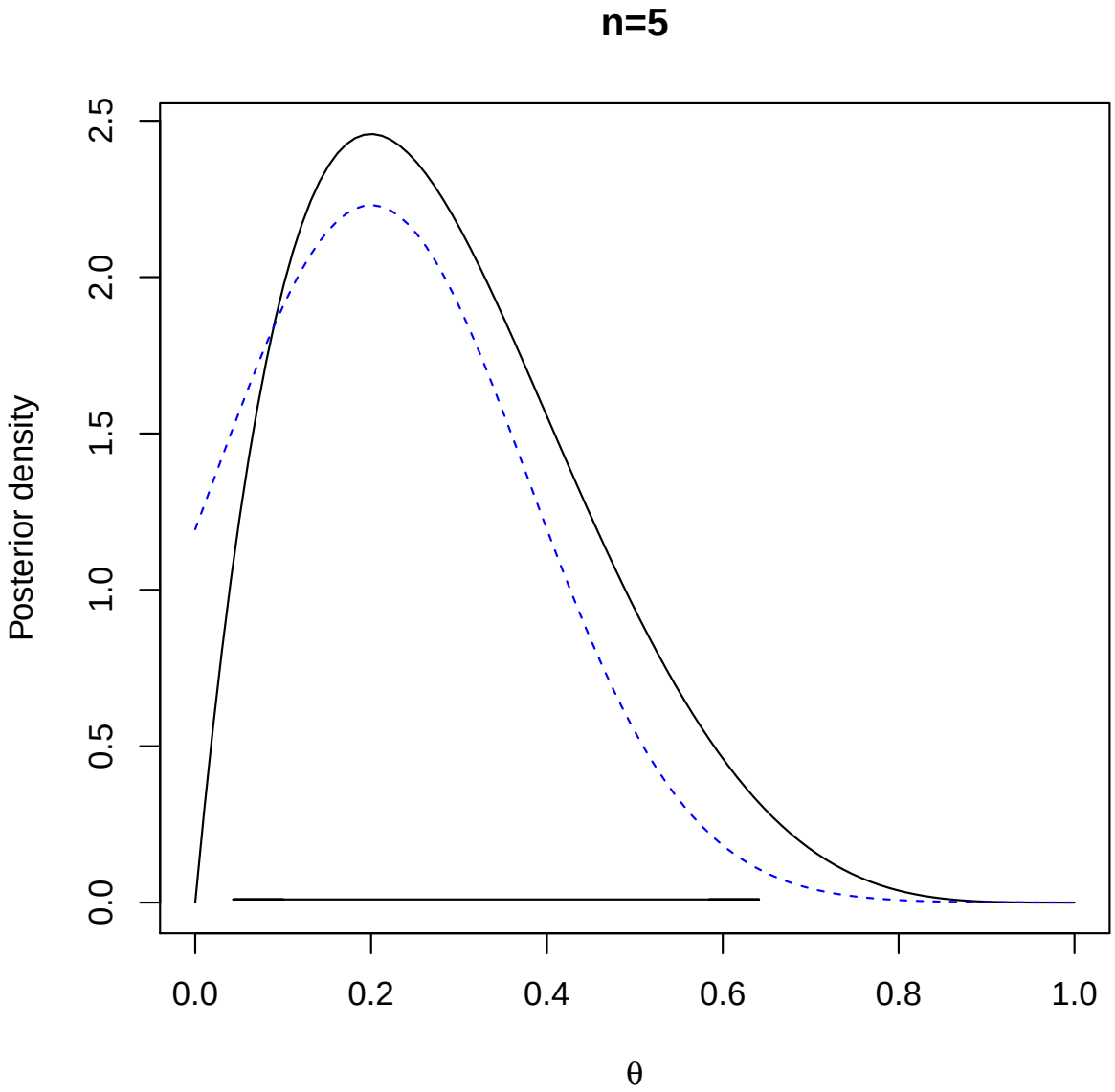
$$y_i \sim \mathbb{B}(\theta)$$

where we are interested in estimating the prob. of success θ. Prior is uniform over θ, so that:

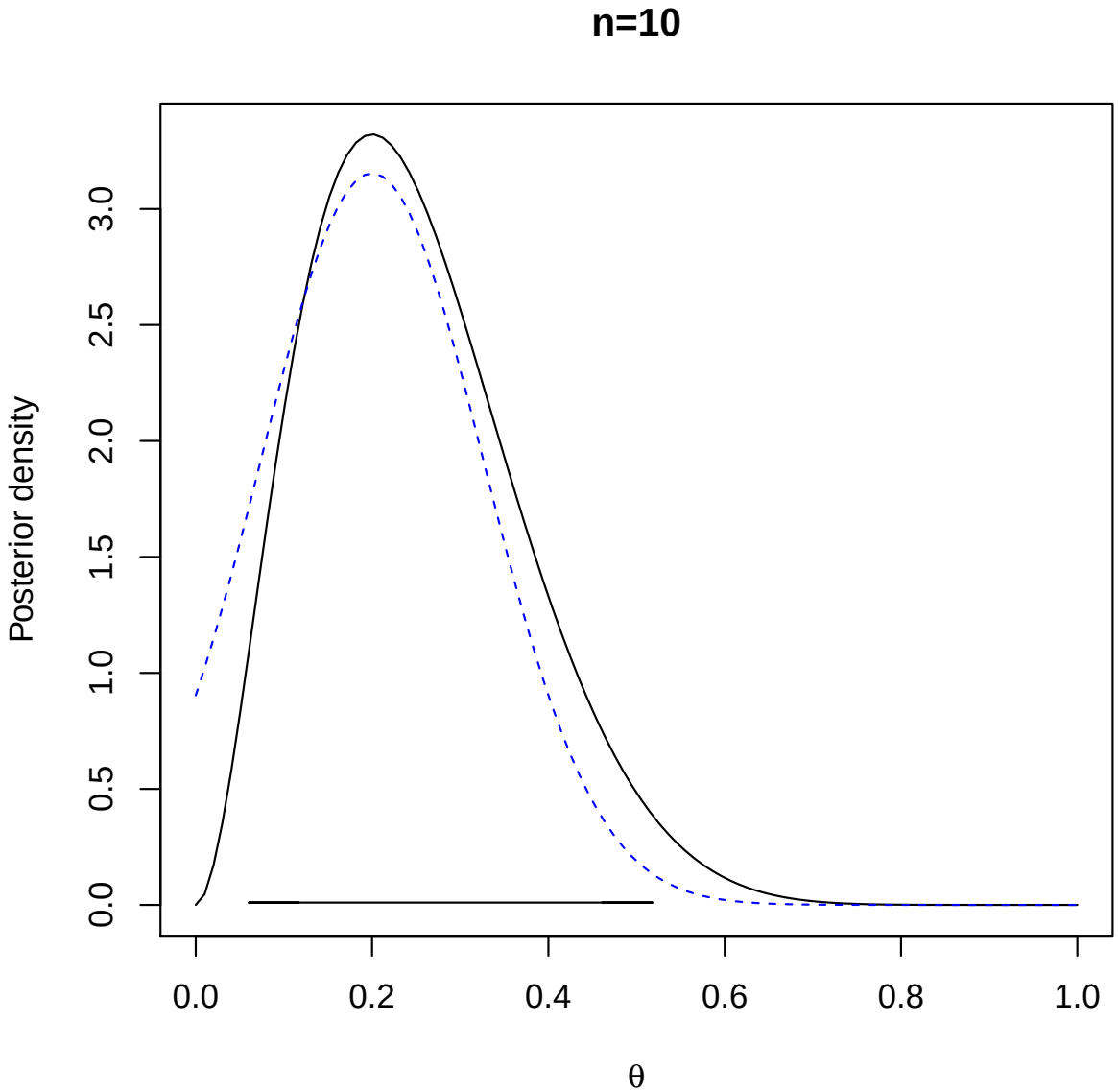
$$p(\theta|y_1 \dots y_n) \propto \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$$

NB: picking a Gaussian approximation for this posterior is rather silly.

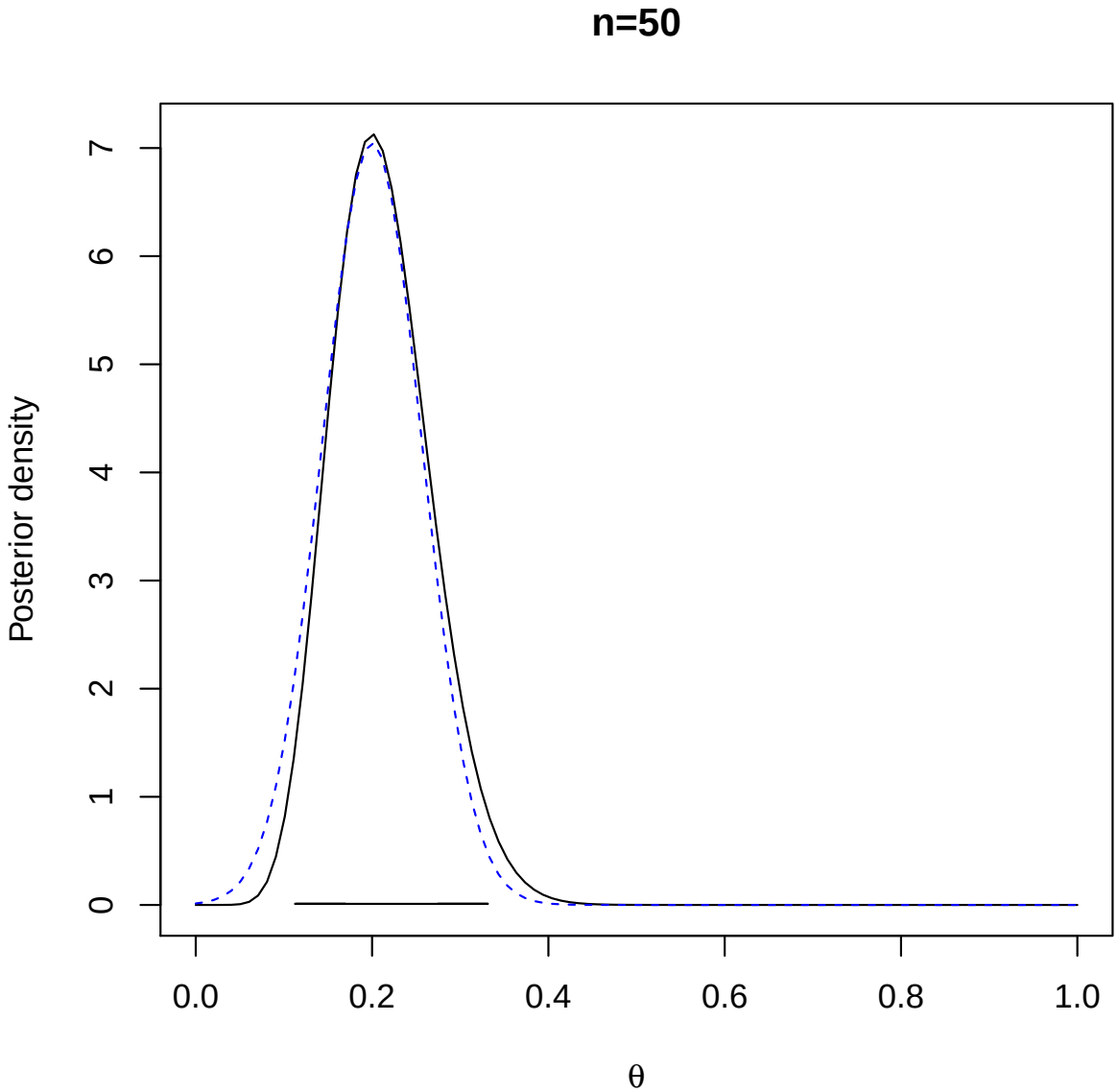
The CGA on the Bernoulli example



The CGA on the Bernoulli example



The CGA on the Bernoulli example



MCMC vs Variational Methods

MCMC pros:

- Modern algorithms are nearly black-box for a lot of models (meaning you run Stan and it works)
- The longer you wait, the better the answer

MCMC cons:

- Slow

MCMC vs Variational Methods

Variational pros:

- Really fast
- Can be very accurate in certain models

Cons:

- Need model-specific work, except for CGA
- Weak theoretical guarantees
- Only captures certain features of your posterior (*)
- You won't get a better answer if you wait longer (*)

What are corrections good for?

Corrections try to improve on the last two points

- You can get more out of your approximation
- You can get a better answer if you invest more time

When is it worth improving an approximation?

- At this stage, we have an approximation and the goal of the talk is to give you ways of improving it. These improvements aren't free. Should we bother?
- We are all familiar with the idea that imprecise quantities needn't be reported with 15 decimals of precision (e.g. "the probability that it rains tomorrow is 0.567891354").
- In the same fashion, if you're estimating the distance to the sun, and the posterior density has a standard deviation of, say 10km, nobody cares if your approximation to the mean is off by a cm or two.

When is it worth improving an approximation?

- Suggestion: let's start by estimating an order of magnitude for the errors of the CGA, to see what we could gain with a bit more effort.
- The CGA is exact in large n (lots of data), so we'll estimate the error in the approximation as a function of n.

How good is the CGA?

The CGA gives an approximate mean and variance, and it's a Gaussian. We can divide the question of accuracy into two parts:

1. How good are the approximate mean and variance?
2. How good can a Gaussian approximation to π be, even with the right mean and variance?

How good is the approximate mean?

The CGA replaces the mean with the mode. For unimodal distributions (Basu & DasGupta, 1997):

$$|E(\theta) - \theta^{\star}| \leq \sqrt{\frac{3}{5} Var(\theta)} \tag{1}$$

so we know the error is at most $\mathcal{O}\left(Var(\theta)^{1/2}\right)$.

How good is the approximate mean?

- Typically, $Var(\theta) = \mathcal{O}(n^{-1})$
- Implies that the variational error is (mode - mean) $\mathcal{O}(n^{-1/2})$
- That's the worst possible rate for unimodal distribution, sometimes the CGA is much better than that

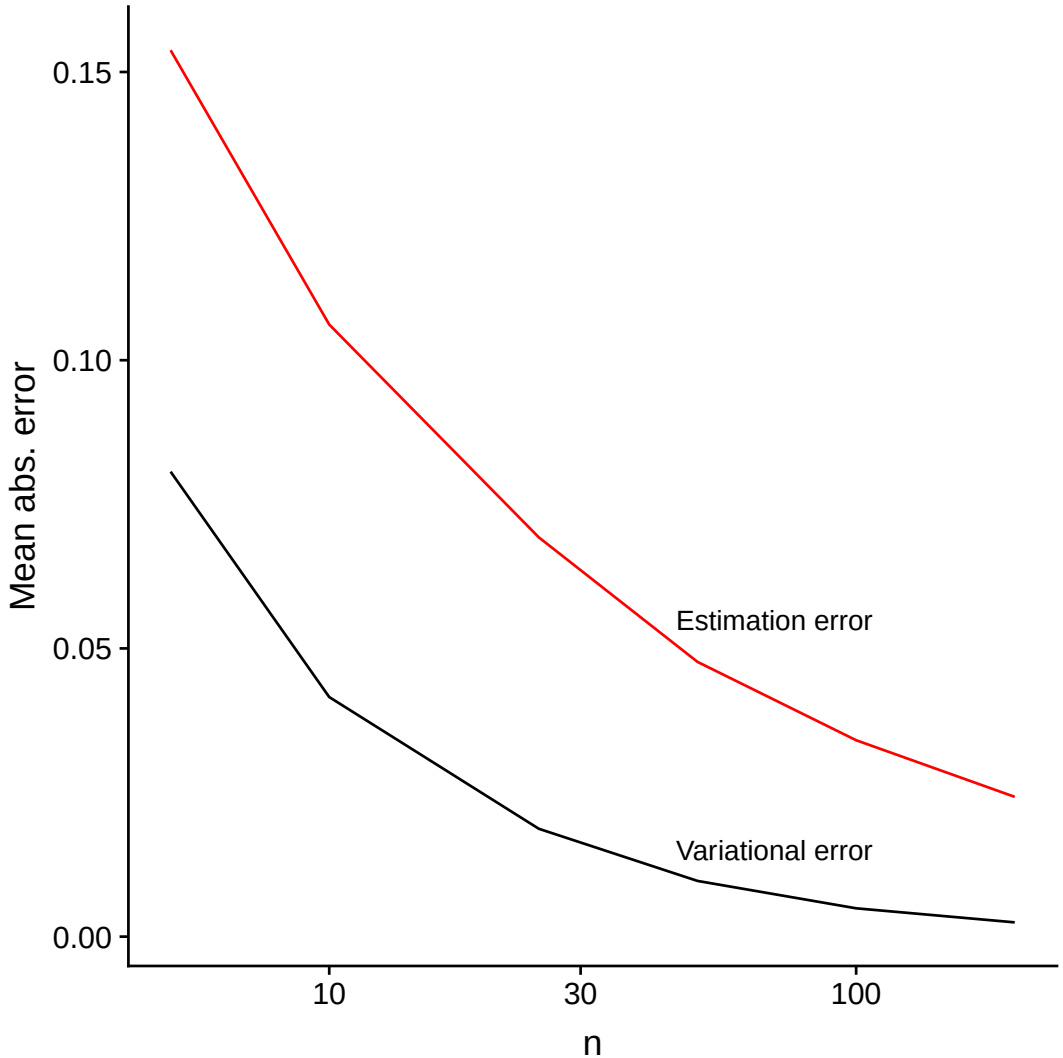
How good is the approximate variance?

Under the hypothesis that π is strongly log-concave and some additional assumptions (Dehaene & Barthelme 2015), we have a *relative* error of $\mathcal{O}(n^{-1})$ for the variance, meaning $\mathcal{O}(n^{-1/2})$ for the standard deviation.

What does that tell us?

- Standard asymptotics: distance between "true" θ and the posterior mean $E(\theta|y_1 \dots y_n)$ is $\mathcal{O}(n^{-1/2})$.
- In some cases, the error in the CGA may have the same magnitude as the estimation error
- That's going to be the case in posterior distributions with a lot of skew

Estimation error vs. variational error



Bernoulli example, $\theta = 0.25$

Conclusion so far

- For small n and/or hard posteriors, you can improve the CGA by a significant margin
- For large n and/or easy posteriors, not worth the bother
- Of course no one knows in advance if they are dealing with an easy or a hard case (Open Research Problem #1)

The Laplace approximation

- We need to introduce the Laplace approximation proper.
- Let $Z = \int p(\theta|y)d\theta = \int \pi(\theta)d\theta$ (marginal likelihood).
- Laplace approximation is closely related to the CGA. Note that the approximation is $q(\theta) = \frac{1}{\sqrt{2\pi h^{-1}}} \exp(-\frac{h}{2}(\theta - \theta^\star)^2)$
- If $q(\theta) \approx \pi(\theta)/Z$ then $Z \approx \pi(\theta)/q(\theta)$
- In particular, at $\theta^\star$ we obtain:

$$Z \approx \sqrt{2\pi h^{-1}}\pi(\theta^\star)$$

- That's the Laplace approximation proper

Accuracy of the Laplace approximation

Tierney & Kadane (1986): the relative error in the Laplace approximation is $\mathcal{O}(1/n)$, i.e.

$$Z = Z_{lap}(1 + \mathcal{O}(n^{-1}))$$

Note that this is definitely better than what we had for the mean and std. dev. We will come back to that later.

Summary on Laplace approximation

- Laplace approximation is an approximation to integrals (Z), not distributions
- Closely related to the CGA
- It's quite reliable ($\mathcal{O}(1/n)$)

Improvement via importance sampling

- This first idea is obvious: if you have an approximation, sample from it, then use importance sampling to compute expectations
- Applying this idea to the computation of Z (marginal likelihood):

$$Z = \int \pi(\theta) d\theta = \int q(\theta) \frac{\pi(\theta)}{q(\theta)} d\theta \approx (1/m) \sum_{i=1}^{m} \frac{\pi(\theta_i)}{q(\theta_i)}$$

for $\theta_1 \ldots \theta_m$ drawn IID from q. Call $\tilde{Z}_m$ this approximation.

Importance sampling: estimating the mean

Along the same lines, the obvious IS estimator for the mean is the following:

$$E(\theta) = Z^{-1} \int \pi(\theta)\theta d\theta = Z^{-1} \int q(\theta)\theta \frac{\pi(\theta)}{q(\theta)} d\theta \approx \tilde{Z}_m^{-1} \sum_{i=1}^{m} \theta_i \frac{\pi(\theta_i)}{q(\theta_i)}$$

ie., the "self-normalised" importance sampling estimate.

How well can we expect IS to do?

Assume the variational approximation is perfect, i.e. $q(\theta) = Z^{-1}\pi(\theta)$. Then:

$$\tilde{Z}_m = (1/m)\sum_{i=1}^{m} \frac{\pi(\theta_i)}{q(\theta_i)} = (1/m)Z = Z$$

We get a zero-variance estimator!

How well can we expect IS to do?

IS estimate for the mean is just the empiral mean of the sample $\theta_1, \ldots, \theta_m$.
In other words, we started out with a perfect estimate, and all we have done is add Monte Carlo error (of order $\mathcal{O}(m^{-1/2})$)

How well can we expect IS to do?

More generally, if we assume that $q(\theta)$ is "not far" from π, in the sense that:

$$q(\theta) \propto \pi(\theta) \exp(\epsilon d(\theta))$$

for small ϵ, we get a $\mathcal{O}(\epsilon m^{-1/2})$ relative error for Z and a $\mathcal{O}((1+\epsilon)m^{-1/2})$ error for the mean.
If the variational approximation is very bad (q far from π), IS estimators can be arbitrarily awful and it's hard to say much.

Summary

- Importance Sampling has the nice property of being convergent with greater effort (the more computational effort you put in, the better the estimate, on average).
- It has the less nice property that if your approximation is already exact, IS will make it worse.
- So what else can we do?

Correction via perturbation

- The next class of methods we will look at are based on perturbations.
- These ideas have been used forever in statistical physics,
- In Bayesian statistics they also appear all over the place.
- Essential principle: if you can't compute a moment, compute a derivative (and vice-versa).

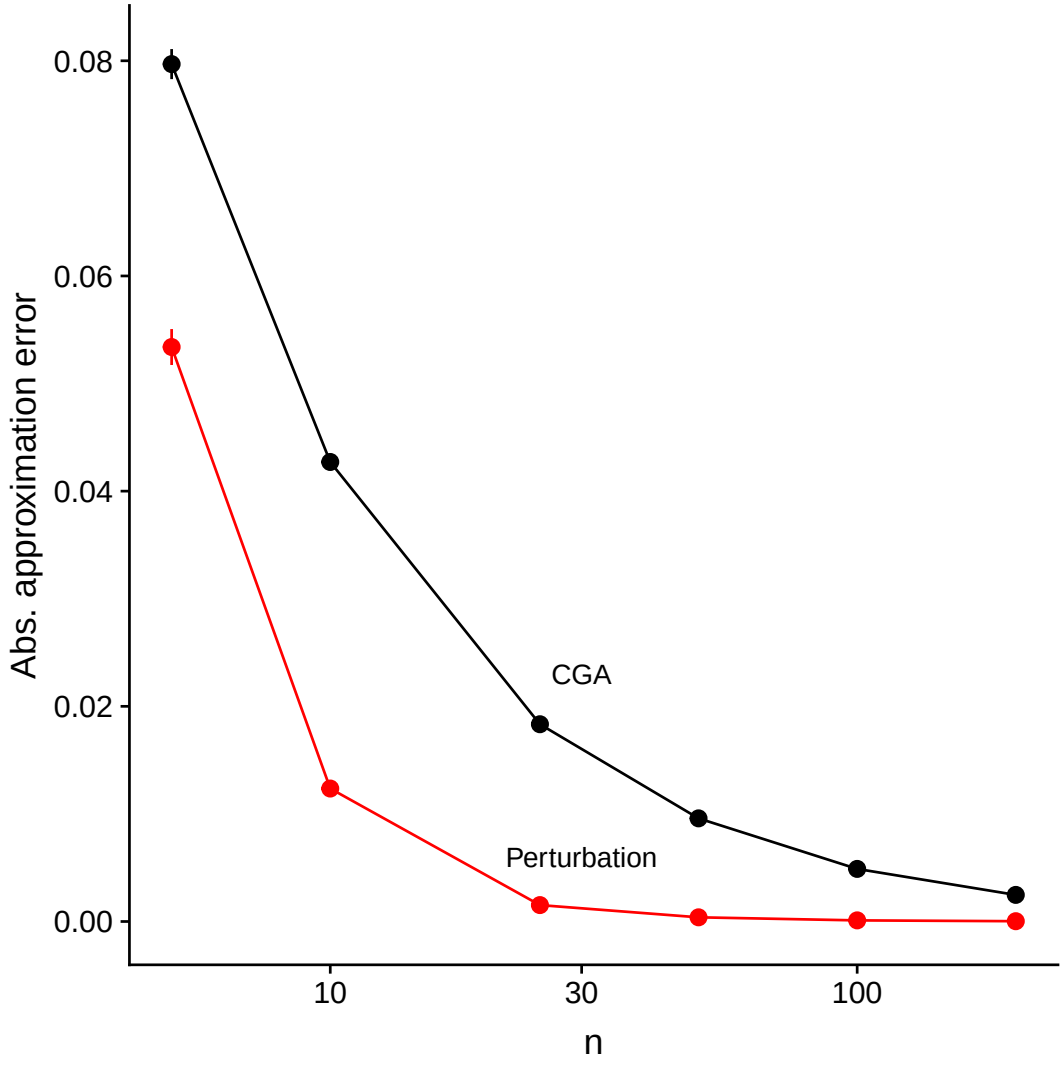
The cumulant generating function

- Let: $\Psi(t) = \log \int \pi(\theta) \exp(t\theta) d\theta$
- Cumulant generating function of Ψ
- Easy to check that $\frac{\partial}{\partial t}\Psi(t)|_{t=0} = \frac{\int \theta\pi(\theta)d\theta}{\int \pi(\theta)d\theta}$, i.e. the mean of π

The CGF as a perturbation

- Interpretation: given an algorithm that outputs an approximation to $\log Z$ (the integration constant), we can construct an algorithm that outputs an approximation to the mean of π
- Step 1: run the algorithm a first time with π, to get $\log \tilde{Z} \approx \log \int \pi(\theta) d\theta$
- Step 2: run the algorithm again, on a small perturbation of π, specifically: $\pi_\epsilon(\theta) = \pi(\theta) \exp(\epsilon\theta)$. The algorithm outputs $\log \tilde{Z}_\epsilon \approx \log \int \pi_\epsilon(\theta) d\theta$
- The approximate mean is $\frac{\log \tilde{Z}_\epsilon - \log \tilde{Z}}{\epsilon}$

Results

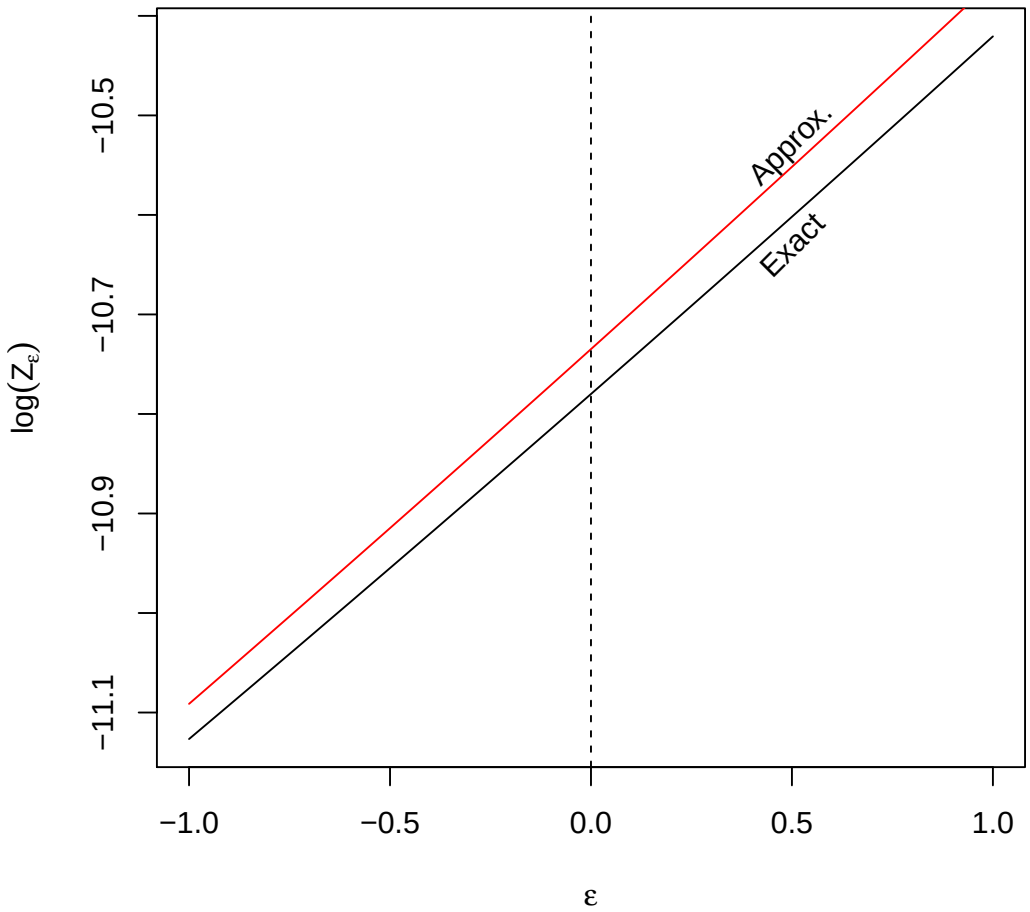


Results on the Bernoulli example ($\theta = 0.25$)

The cancellation miracle

- Why does this work so much better than the CGA?
- Part of the error cancels, and a quick picture will help understand why

The cancellation miracle



Generalising

- It's easy to generalise to other moments of π: if you need to compute $E_\pi(g(\theta))$, use the following perturbation

$$\log Z(\epsilon) = \log \int \pi(\theta) \exp(\epsilon g(\theta)) d\theta$$

- Using $\log(1+x) = x + \mathcal{O}(x^2)$ and $\exp(x) = 1 + x + \mathcal{O}(x^2)$ we find

$$\begin{aligned} \log Z(\epsilon) &= \log \int \pi(\theta)(1 + \epsilon g(\theta) + \mathcal{O}(\epsilon^2)) d\theta \\ &= \log Z + \epsilon E_\pi(g) + \mathcal{O}(\epsilon^2) \end{aligned}$$

Corrections for the Laplace expansion

- In the case of the Laplace expansion $\frac{d}{d\epsilon} \log \tilde{Z}(\epsilon)$ can be carried out analytically
- We find:

$$\frac{d}{d\epsilon} \log \tilde{Z}(\epsilon) = g(\theta^\star) - \frac{g'(\theta^\star)\psi'''(\theta^\star)}{2(\psi''(\theta^\star))^2}$$

- This only depends of $g(\theta^\star)$ and $g'(\theta^\star)$, so the correction is extremely local
- Note that $E_\pi(g) \approx g(\theta^\star)$ is the naive approximation we'd get from the CGA (neglecting high order derivatives of g).
- The correction relative to the naive approximation has order $\mathcal{O}(n^{-1})$.

Some notes on implementation

- You may be able to carry out the differentiation analytically, however painful that sounds (it's possible for EP, for example).
- You may want to use centred differences rather than forward differences: $(1/2\epsilon)\left(\log Z(\tilde{\epsilon}) - \log Z(\tilde{-\epsilon})\right)$
- An obvious alternative is to use automatic differentiation but I haven't tried it.

Extension: computing covariances

- A very cheap class of variational approximations use fully-factorising distributions, i.e. $q(\theta_1 \dots \theta_d) \propto q_1(\theta_1) \dots q_d(\theta_d)$
- The approximate posterior distribution is independent over parameters, and so carries no information on the covariance.
- A similar perturbation trick can be used, however.

Extension: computing covariances

- Idea: fully-factorised q still gives an approximation to the mean and to $\log Z$, and we may approximate

$$\boldsymbol{\kappa}(\boldsymbol{\epsilon}) = \int \boldsymbol{\theta} \exp(\boldsymbol{\epsilon}^t \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

- The rest is left as an exercise to the reader

Conclusion

- A viable alternative to MCMC in Bayesian statistics is a combination of decent variational method + correction where needed
- Sometimes a CGA is all you really need, and it has the advantage of requiring very little model-specific work

More open problems

- Natural heuristic: if the correction is small, that means the initial approximation was good
- Open Problem # 2: prove that is not just a heuristic
- Open Problem # 3: find explicit error bounds

Bonus material: Paquet-Winther-Opper corrections

- Neat technique described in Paquet, Winther & Opper (2009): assume the posterior can be written as

$$\pi(\theta) = \prod_{i=1}^{n} p(y_i|\theta)p(\theta)$$

- Similarly, assume that the approximation can be written as:

$$q(\theta) = \prod_{i=1}^{n} q_i(\theta)p(\theta)$$

where each $q_i(\theta)$ approximates $p(y_i|\theta)$, so that

$$\frac{p(y_i|\theta)}{q_i(\theta)} = 1 + \epsilon_i(\theta)$$

and ϵ_i is small.

Bonus material: Paquet-Winther-Opper corrections

Then:

$$\pi(\theta) = \pi(\theta)\frac{q(\theta)}{\pi(\theta)} = \prod_{i=1}^{n}(1 + \epsilon_i(\theta))$$

Expanding to first order in ϵ:

$$\begin{aligned}\pi(\theta) &\approx q(\theta)\left(1 + \sum_{i=1}^{n}\epsilon_i(\theta) + \mathcal{O}(\epsilon_i\epsilon_j)\right) \\ &= q(\theta) + \sum_i\left(q(\theta)\frac{p(y_i|\theta)}{q_i(\theta)} - 1\right)\end{aligned}$$

PWO corrections

Terms like: $q(\theta)\frac{p(y_i|\theta)}{q_i(\theta)}$ can be understood as "remove approximation for the i-th factor and replace with the true factor instead".
If you apply the PWO correction to the CGA you get something that's very close to Expectation Propagation.

References

Some references from physics:

Opper, M., & Winther, O. (2001). From naive mean field theory to the TAP equations. Advanced mean field methods: theory and practice, 7-20.

Mezard, M., & Montanari, A. (2009). Information, physics, and computation. Oxford University Press.

From the Bayesian literature:

Tierney, L., Kass, R. E., & Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. Journal of the American Statistical Association, 84(407), 710-716.

Opper, M., Paquet, U., & Winther, O. (2013). Perturbative corrections for approximate inference in

Gaussian latent variable models. The Journal of Machine Learning Research, 14(1), 2857-2898.

For error bounds, see:

Dehaene, G. P., & Barthelmé, S. (2015). Bounding errors of expectation-propagation. In Advances in

Neural Information Processing Systems (pp. 244-252).