STAT G8325
Gaussian Processes and Kernel Methods
Lecture Notes §03: Approximate Inference

John P. Cunningham

Department of Statistics
Columbia University

Outline

Administrative interlude

Non-conjugate gp models

Basic approaches: Laplace and EP [KR05]

Elliptical slice sampling [MAM09]

Integrated nested Laplace approximations [RMC09]

Outline

Administrative interlude

Non-conjugate gp models

Basic approaches: Laplace and EP [KR05]

Elliptical slice sampling [MAM09]

Integrated nested Laplace approximations [RMC09]

Progress...

Week	Lectures	Content
1	Sep 9	Introduction to gaussian processes for machine learning
2	Sep 14,16,21,23	Model selection
3	Sep 23,28,30	Approximate inference
		• Reading: [KR05]; [RMC09]; [MAM09]; [RW06, ch. 3; 5.5]
4	Oct 5,7	Kernels
5	Oct 12,14	Speed and scaling part 1: reduced-rank processes

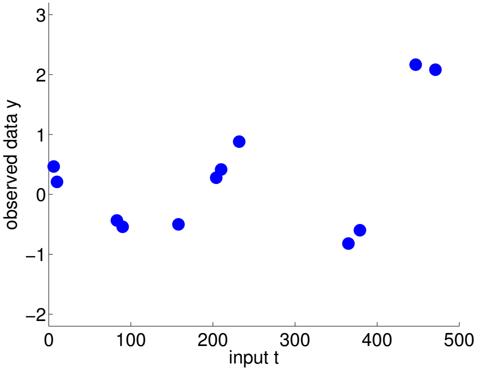
Outline

Approximate inference in words

Structured regression models are perhaps the most commonly used class of models in statistical applications. We consider approximate Bayesian inference in ... latent Gaussian models, where the latent field is Gaussian, controlled by a few hyperparameters and with non-Gaussian response variables. The posterior marginals are not available in closed form owing to the non-Gaussian response variables [RMC09].

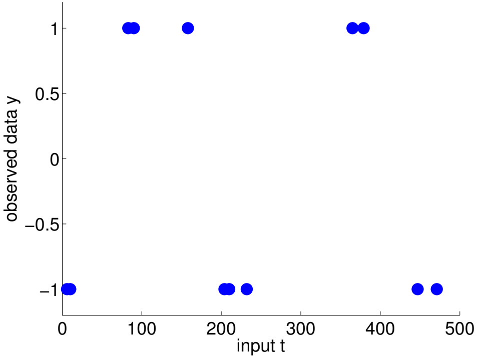
Continuous obervations

- Thus far we have considered gaussian observations
- (where continuous regression made sense)
- Our likelihood model was $y_i | f_i \sim \mathcal{N}(f_i, \sigma_\epsilon^2)$



Binary label data

- Now consider the classification setting for $y_i \in \{-1, +1\}$.
- $y_i | f_i \sim \mathcal{N}(f_i, \sigma_\epsilon^2)$ is inappropriate.

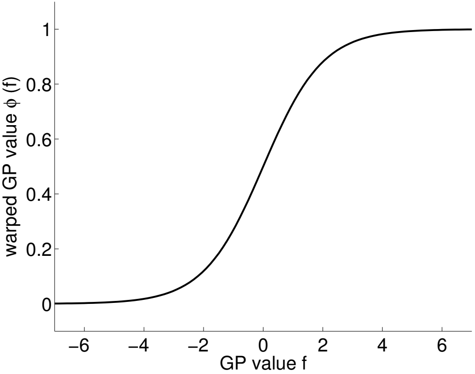


Classification with gp

- Probit or Logistic "regression" model on $y_i \in \{-1, +1\}$:

$$p(y_i|f_i) = \phi(y_i f_i) = \frac{1}{1 + \exp(-y_i f_i)} \quad \text{or} \quad = \int_{-\infty}^{y_i f_i} \mathcal{N}(u; 0, 1) du.$$

- Warps f onto the $[0, 1]$ interval:

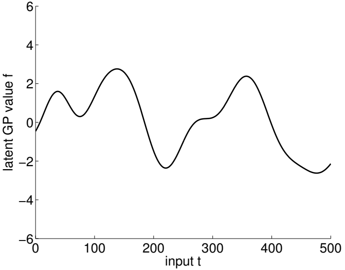


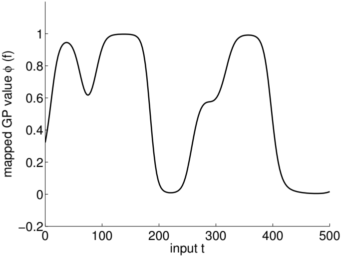
GP Classification

- Probit or Logistic "regression" model on $y_i \in \{-1, +1\}$:

$$p(y_i|f_i) = \phi(y_i f_i) = \frac{1}{1 + \exp(-y_i f_i)} \quad \text{or} \quad = \int_{-\infty}^{y_i f_i} \mathcal{N}(u; 0, 1) du.$$

- Warps f onto the $[0, 1]$ interval:





What we want to calculate our usual quantities

- predictive distribution:

$$p(y^*|y) = \int p(y^*|f^*)p(f^*|y)df^*$$

- predictive posterior:

$$p(f^*|y) = \int p(f^*|f)p(f|y)df$$

- data posterior:

$$p(f|y) = \frac{\prod_i p(y_i|f_i)p(f)}{p(y)}$$

- *None* of which is tractable to compute.

Structured approximate inference

- predictive distribution:

$$p(y^*|y) = \int p(y^*|f^*)q(f^*|y)df^*$$

- predictive posterior:

$$q(f^*|y) = \int p(f^*|f)q(f|y)df$$

- data posterior:

$$q(f|y) \approx p(f|y) = \frac{\prod_i p(y_i|f_i)p(f)}{p(y)}$$

- Structured (gaussian) approximations: [KR05]; [RMC09]; [RW06, ch. 3; 5.5].

Approximate inference via sampling

- predictive distribution:

$$p(y^*|y) = \int p(y^*|f^*)q(f^*|y)df^*$$

- predictive posterior:

$$q(f^*|y) = \int p(f^*|f)q(f|y)df$$

- data posterior:

$$q(f|y) = \mathcal{N}(m_s, K_s), \qquad m_s = \frac{1}{m}\sum_{k=1}^{m} f_k, \quad K_s = ..., \qquad f_k \sim p(f|y).$$

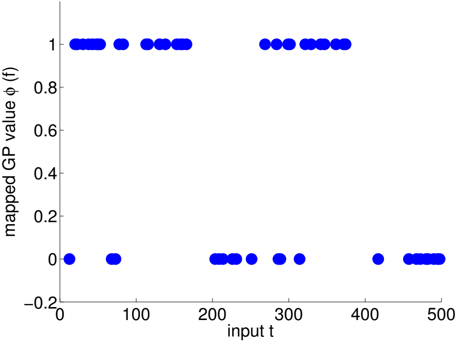
- Sampling approximations: [MAM09].

Approximate inference in a nutshell

- Methods for producing a (typically gaussian) $q(f|y) \approx p(f|y)$.
- Laplace approximation, expectation propagation, sampling, variational inference, others (probably in that order, though changing).
- Remember these are technologies within a gp method.
- Subject of much research and often work well.

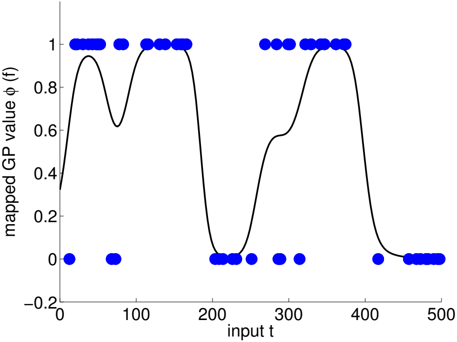
Using approximate inference

- Allows "regression" on the $[0, 1]$ interval



Using approximate inference

- Allows "regression" on the $[0, 1]$ interval



Outline

Laplace approximation

- Fit a gaussian $q(f) = \mathcal{N}(f; \hat{f}, A)$ to the posterior as:

$$\begin{aligned}\log p(f|D,\theta) \propto \Psi(f) &= \log p(D|f) - \frac{1}{2}\log|K_\theta| - \frac{1}{2} f^\top K_\theta^{-1} f - \frac{n}{2}\log 2\pi \\ &\approx \log q(\hat{f}) - \frac{1}{2}(f-\hat{f})^\top A^{-1}(f-\hat{f}).\end{aligned}$$

- $\hat{f} = \arg\max_f \Psi(f)$ is the posterior mode, and curvature $A = -(\nabla^2_{\hat{f}}\Psi)^{-1}$.

Why negative?

- q is implied by the second order Taylor expansion of the log posterior.
- Depends on $\theta \rightarrow$ workable for model selection (ML-II, etc.).

What changes with θ?

- This approximation is usually easiest, and can be further simplified.

Typical conveniences in Laplace approximations

- Log concavity in $p(D|f) \to$ finding $\hat{f}$ is a convex problem.
- Newton's method then has simple iterates:

$$f^{k+1} = f^k - \alpha_k (\nabla^2 \Psi)^{-1} \nabla \Psi.$$

- Conditional independence $p(D|f) = \prod_{i=1}^n p(y_i|f_i)$ implies:

$$\nabla_f^2 \Psi = K_\theta^{-1} + D_\theta \qquad \text{for diagonal} \quad D = \left\{ \frac{d^2}{df_i^2} p(y_i|f_i) \right\}_{ii}.$$

- Then use the matrix inversion lemma for efficiency:

$$A = -(\nabla_{\hat{f}}^2 \Psi)^{-1} = (K^{-1} + D)^{-1} = K - K(D^{-1} + K)^{-1} K.$$

Marginal likelihood in Laplace approximations

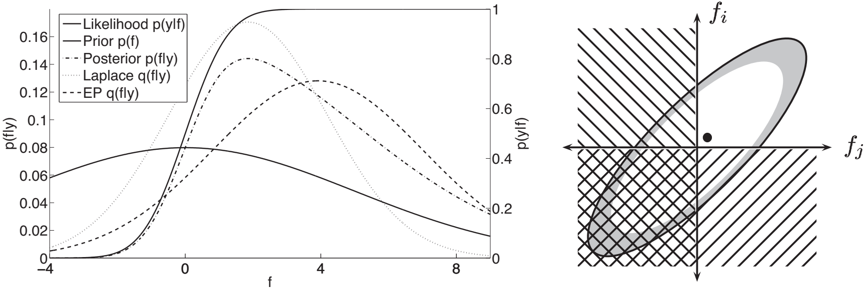
- Marginal likelihood:

$$\begin{aligned}\log p(D|\theta) &= \log \int p(D|f)p(f|\theta)df = \log \int \exp\left\{\Psi(f)\right\} df \\ &\approx \log \int \exp\left\{\log q(\hat{f}) - \frac{1}{2}(f-\hat{f})^\top A^{-1}(f-\hat{f})\right\} df \\ &= \log q(\hat{f}) + \frac{1}{2}\log|A| + \frac{n}{2}\log 2\pi.\end{aligned}$$

- Laplace is a local gaussian approximation that enables all desired operations.
- ...but how well?
- The field (and [KR05] in particular) use gp classification as the benchmark.

Discuss the polya-gamma augmentation trick next week?

Visualizing a Laplace approximation [KR05, fig. 1]



(ignore EP for now...)

Expectation propagation

- Warning: EP is often empirically better, theoretically (far) less pleasing.
- Again a gaussian approximation (though it can be more general):

$$p(f|D) \quad = \quad \frac{p(f)\prod_{i=1}^{n} p(y_i|f_i)}{p(D)} \quad \approx \quad \frac{p(f)\prod_{i=1}^{n} t(y_i|f_i)}{q(D)} \quad \triangleq \quad q(f|D),$$

- where $t(y_i|f_i) = Z_i \mathcal{N}(f_i; \mu_i, \sigma_i^2)$ and $q(f|D) = \mathcal{N}(f; m, A)$.
- Note θ has been suppressed, but model selection again available via $q(D)$.
- Here as in many places we assume wlog that the gp $p(f)$ has zero mean.
- Letting $\Sigma = \text{diag}\left(\sigma_1^2, ..., \sigma_n^2\right)$, the posterior q then has

$$m = A\Sigma^{-1}\begin{bmatrix}\mu_1 \\ \vdots \\ \mu_n\end{bmatrix}, \qquad A = \left(K^{-1} + \Sigma^{-1}\right)^{-1}.$$

- We now must choose Z_i, μ_i, and σ_i^2...

EP key idea: cavity and tilted distributions

- Finding a global approximation $q(f)$ is hard $\rightarrow$ iterate locally.
- For $i \leftarrow 1$ to n
  - Cavity: remove site approximation $t(f_i)$ from $q(f)$ to form $q_{\setminus i}(f)$.
  - Tilt: add true site to form tilted $q_{\setminus i}(f)p(y_i|f_i)$.
  - Approximate: $\arg\min \; D_{KL}\left(q_{\setminus i}(f)p(y_i|f_i)||q_{\setminus i}(f)t(f_i)\right)$.

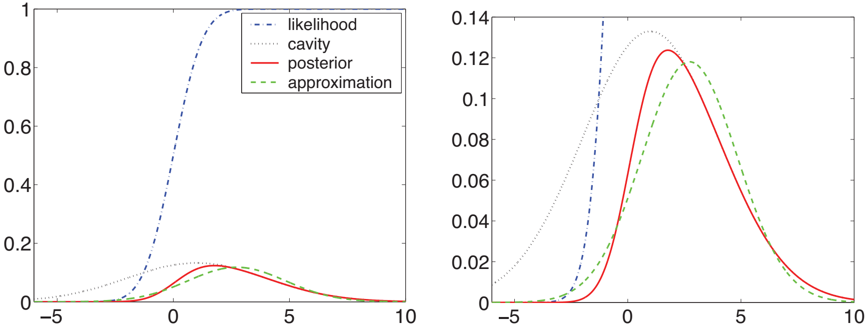
minimize over what?

what form is the cavity?

- Repeat above loop until convergence.
- Closure of exponential family under division/multiplication $\rightarrow$ subtraction/addition of natural parameters.
- Minimizing $D_{KL}\left(p||q\right)$ for normal q is moment matching.

everyone comfortable with that claim?

What EP is doing [RW06, fig. 3.4]



Moment matching in EP

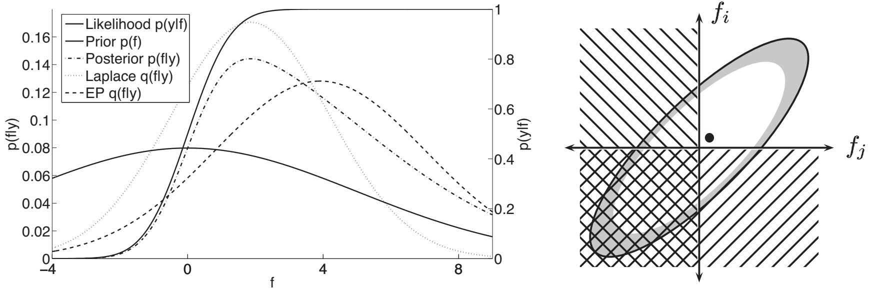
- $\arg\min \; D_{KL}\left(q_{\setminus i}(f)p(y_i|f_i)||q_{\setminus i}(f)t_i(f_i)\right)$.
- To match moments, choose $t_i(f_i) = Z_i\mathcal{N}(f_i; \mu_i, \sigma_i^2)$ with:

$$
\begin{aligned}
\sigma_i^2 &\leftarrow \left(\left(m_2 - m_1^2\right)^{-1} - \sigma_{\setminus i}^{-2}\right)^{-1} \\
\mu_i &\leftarrow \sigma_i^2\left(m_1(\sigma_{\setminus i}^{-2} + \sigma_i^{-2}) - \frac{\mu_{\setminus i}}{\sigma_{\setminus i}^2}\right) \\
Z_i &\leftarrow m_0\sqrt{2\pi(\sigma_{\setminus i}^2 + \sigma_i^2)}\exp\left\{\frac{(\mu_i - \mu_{\setminus i})^2}{2(\sigma_{\setminus i}^2 + \sigma_i^2)}\right\}.
\end{aligned}
$$

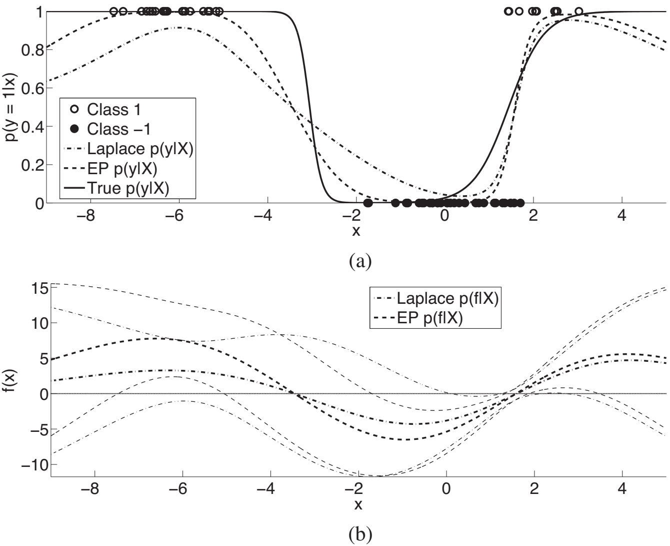
- where m_0, m_1, m_2 are the moments of $q_{\setminus i}(f)p(y_i|f_i)$.
- Choose site approximation to match moments of the tilted distribution.
- Subsequent approximation of the normalizer:

$$
\begin{aligned}
\log p(D) \approx \log q(D) &= \log\int q(f)\prod_{i=1}^{n} t(f_i) \\
&= \sum_{i=1}^{n}\log Z_i - \frac{1}{2}\log|K+\Sigma| - \frac{1}{2}\mu^\top(K+\Sigma)^{-1}\mu - \frac{n}{2}\log 2\pi.
\end{aligned}
$$

Visualizing the EP approximation [KR05, fig. 1]

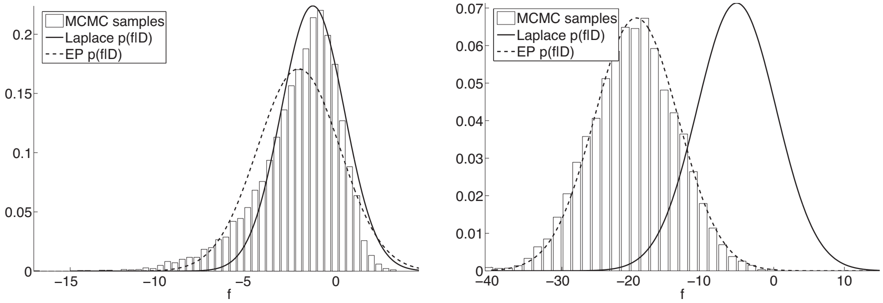


Visualizing EP and Laplace [KR05, fig. 3]



- Note: same as [RW06, fig. 3.6] (as much of this paper is).

Visualizing EP and Laplace [KR05, fig. 7]



(a) (b)

- MCMC here is MH variant (HMC).
- AIS used when comparing normalizing constant approximations (see paper).

Quantifying EP and Laplace [KR05, table 1]

			Laplace			EP			SVM	
Data Set	n	d	E	I	$\|\|\mathbf{m}\|\|$	E	I	$\|\|\mathbf{m}\|\|$	E	I
Ionosphere	351	34	8.84	0.591	49.96	7.99	0.661	124.94	5.69	0.681
Wisconsin	683	9	3.21	0.804	62.62	3.21	0.805	84.95	3.21	0.795
Pima Indians	768	8	22.77	0.252	29.05	22.63	0.253	47.49	23.01	0.232
Crabs	200	7	2.0	0.682	112.34	2.0	0.908	2552.97	2.0	0.047
Sonar	208	60	15.36	0.439	26.86	13.85	0.537	15678.55	11.14	0.567
USPS 3 vs 5	1540	256	2.27	0.849	163.05	2.21	0.902	22011.70	2.01	0.918

- Let's question the worth of bayesian approaches here...

Takeaways

- Using gp in non-conjugate settings is common.
- Approximate inference machinery is required and is nontrivial.
- Laplace and EP seem to carry the usual performance/complexity tradeoff.
- Classification with gp has become a standard test benchmark.
- Note: numerous computational tricks are available; see [KR05, RW06].

Outline

What we want to calculate our usual quantities

- predictive distribution:

$$p(y^*|y) = \int p(y^*|f^*)p(f^*|y)df^*$$

- predictive posterior:

$$p(f^*|y) = \int p(f^*|f)p(f|y)df$$

- data posterior:

$$p(f|y) = \frac{\prod_i p(y_i|f_i)p(f)}{p(y)}$$

- *None* of which is tractable to compute.

Approximate inference via sampling

- predictive distribution:

$$p(y^*|y) = \int p(y^*|f^*)q(f^*|y)df^*$$

- predictive posterior:

$$q(f^*|y) = \int p(f^*|f)q(f|y)df$$

- data posterior:

$$q(f|y) = \mathcal{N}(m_s, K_s), \qquad m_s = \frac{1}{m}\sum_{k=1}^{m} f_k, \quad K_s = ..., \qquad f_k \sim p(f|y).$$

- Sampling approximations: [MAM09].

Elliptical slice sampling

- Bespoke slice sampler for latent Gaussian models
- The go-to tool for sampling gp with complicated non-conjugate likelihoods
- No free parameters enables out-of-the-box use
- Simple code in matlab (not in gpy)

  excellent project... see gpy/inference/mcmc

- Still a sampler... possibly quite slow.

Starting point: MH from [Nea98]

- General non-conjugate posterior form:

$$p(f|D) = \frac{1}{p(D)} p(D|f) p(f) \triangleq \frac{1}{Z} L(f) \mathcal{N}(f; 0, K)$$

- Implement standard Metropolis Hastings MCMC with proposal:

$$q(f'|f) = \sqrt{1-\epsilon^2} f + \epsilon \nu, \qquad \nu \sim \mathcal{N}(0, K)$$

rather out of the blue in [Nea98]

- which will accept with probability defined by the MH rejection kernel:

$$A(f'|f) = \min\left\{1, \frac{L(f')}{L(f)}\right\} = \min\left\{1, \frac{q(f|f') L(f') \mathcal{N}(f'; 0, K)}{q(f'|f) L(f) \mathcal{N}(f; 0, K)}\right\}.$$

- Unfortunately choosing ϵ well is hard and costly.

Elliptical sampling idea

- Reparameterize the proposal:

$$\begin{aligned} q(f'|f) &= \sqrt{1-\epsilon^2} f + \epsilon\nu, \qquad \nu \sim \mathcal{N}(0, K) \\ &= \nu \sin\theta + f\cos\theta, \qquad \nu \sim \mathcal{N}(0, K) \end{aligned}$$

- Idea then is to sample over this ellipse, i.e. $\theta \sim \text{Uniform}[0, 2\pi]$.
- To make a valid MC (correct invariant distribution), rewrite as:

$$\begin{aligned} \nu_0 &\sim \mathcal{N}(0, K) \\ \nu_1 &\sim \mathcal{N}(0, K) \\ \theta &\sim \text{Uniform}[0, 2\pi] \\ f &= \nu_0 \sin\theta + \nu_1 \cos\theta. \end{aligned}$$

You should be able to prove that $f \sim \mathcal{N}(0, K)$.

Elliptical sampling idea

- Accordingly we can sample from:

$$p(\nu_0, \nu_1, \theta | D, f = \nu_0 \sin\theta + \nu_1 \cos\theta) \propto L(f = \nu_0 \sin\theta + \nu_1 \cos\theta)\mathcal{N}(\nu_0; 0, K)\mathcal{N}(\nu_1; 0, K)$$

- via:

$$\begin{aligned}
\theta &\sim \text{Uniform}[0, 2\pi] \\
\nu &\sim \mathcal{N}(0, K) \\
\nu_0 &\leftarrow f \sin\theta + \nu \cos\theta \\
\nu_1 &\leftarrow f \cos\theta - \nu \sin\theta.
\end{aligned}$$

You should be able to derive those steps.

- That sampled θ, ν_0, ν_1 according to f.
- Now slice sample θ', which steps to a new $f' = \nu_0 \sin\theta' + \nu_1 \cos\theta'$.
- This produces a valid algorithm [MAM09, Alg. 1].

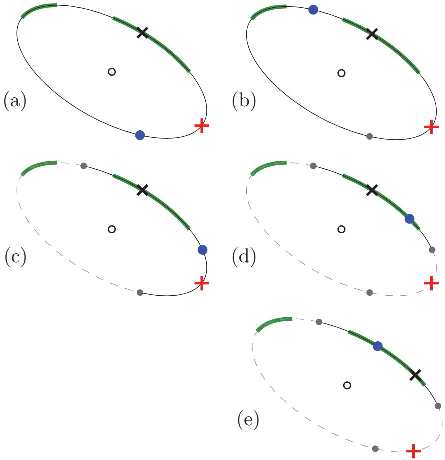
Elliptical slice sampling: [MAM09, Alg. 2]

Algorithm 1 Slice sampling $\theta \to \theta'$

1: Choose ellipse: $\nu \sim \mathcal{N}(0, K)$
2: Likelihood threshold (tolerance):

$$\begin{aligned} u &\sim \text{Uniform}[0,1] \\ \log y &\leftarrow \log L(f) + \log u \end{aligned}$$

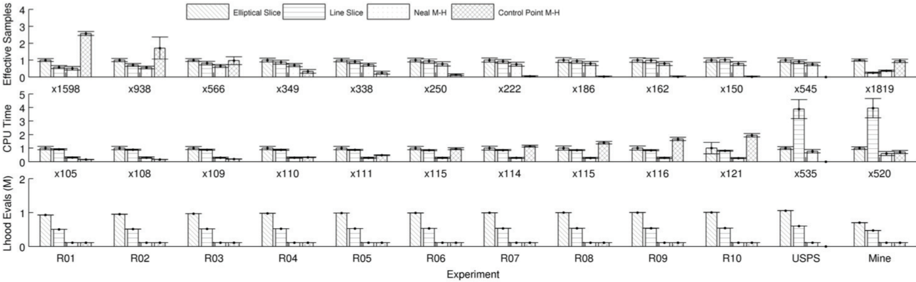
3: Draw proposal $\theta \sim \text{Uniform}[0, 2\pi]$
4: Define slice bracket $[\theta_l, \theta_u] \leftarrow [\theta - 2\pi, \theta]$
5: $f' \leftarrow f \cos\theta + \nu \sin\theta$
6: **if** $\log L(f') > \log y$, **then** return f'
7: **else** change slice interval $\theta_l \leftarrow \theta$ or $\theta_u \leftarrow \theta$
8: Redraw $\theta \sim \text{Uniform}[\theta_l, \theta_u]$
9: Repeat from 5.



- log-likelihood threshold is a tolerance.
- No tuning/extra parameters.
- "Stepping-in" procedure for slice.

Elliptical slice sampling: [MAM09, Alg. 2]

- Producing the results (refer back to §02 and [MA10]):



Outline

Administrative interlude

Non-conjugate gp models

Basic approaches: Laplace and EP [KR05]

Elliptical slice sampling [MAM09]

Integrated nested Laplace approximations [RMC09]

Problems of interest in [RMC09]

- Posterior marginals:

$$
\begin{aligned}
p(f_i|D) &= \int p(f_i|\theta, D) p(\theta|D) d\theta \\
p(\theta_j|D) &= \int p(\theta|D) d\theta_{-j}
\end{aligned}
$$

- [RMC09] uses nested approximations of the form:

$$
\begin{aligned}
q(f_i|D) &= \int q(f_i|\theta, D) q(\theta|D) d\theta \\
q(\theta_j|D) &= \int q(\theta|D) d\theta_{-j}
\end{aligned}
$$

using our notation q, D, f, not $\tilde{\pi}, y, x$ [RMC09, Eq. 2]

- Let's start with $q(\theta|D)$...

Finding the hyperparameter posterior: [RMC09, Eq. 3]

- Let's start with $q(\theta|D)$... a Laplace approximation:

$$q(\theta|D) \quad \propto \quad \frac{p(\hat{f},\theta,D)}{q_{\mathcal{N}}(\hat{f}|\theta,D)}$$

- Why:

$$\begin{aligned} \frac{p(\hat{f},\theta,D)}{q_{\mathcal{N}}(\hat{f}|\theta,D)} &= \frac{p(\hat{f}|\theta,D)p(\theta|D)p(D)}{q_{\mathcal{N}}(\hat{f}|\theta,D)} \\ &\propto \frac{p(\hat{f}|\theta,D)p(\theta|D)}{q_{\mathcal{N}}(\hat{f}|\theta,D)} \\ &\approx p(\theta|D). \end{aligned}$$

- Warning! Not the same Laplace approximation.
- $q_{\mathcal{N}}(f|\theta,D)$ is the previous (and usual) Laplace approximation.
- Previous work suggests $q(\theta|D)$ is particularly accurate...

Editorial interlude: controversial or outdated claims

- *On the other hand, Gaussian approximations are intuitively appealing for latent Gaussian models. For most real problems and data sets, the conditional posterior of f is typically well behaved, and looks 'almost' Gaussian.*

- *The second potential problem is that the iterative process of the basic VB algorithm is tractable for conjugate exponential models only. This implies that $p(\theta)$ must be conjugate with respect to the complete likelihood $p(f, D|\theta)$ and the complete likelihood must belong to an exponential family.*

- Generally [RMC09, §1.6].

- *Rue and Martino (2007) used $q(\theta|D)$ to approximate posterior marginals for θ for various latent Gaussian models. Their conclusion was that $q(\theta|D)$ is particularly accurate: even long MCMC runs could not detect any error in it.*

  (project opportunity)

- Others?

Finding the latent marginal posterior: [RMC09, Eq. 5]

- Now we seek $q(f_i|D)$, the marginal (integrated) posterior of the latent:

$$
\begin{aligned}
p(f_i|D) &= \int p(f_i|\theta, D)p(\theta|D)d\theta \\
&\approx \int q_{\mathcal{N}}(f_i|\theta, D)q(\theta|D) \\
&\approx \sum_{k=1}^{K} q_{\mathcal{N}}(f_i|\theta_k, D)q(\theta_k|D),
\end{aligned}
$$

- where $q_{\mathcal{N}}$ is the marginal of the previous (usual) Laplace approximation,
- and where samples θ_k are drawn according to [RMC09, §3.1].
- Hence, integrated nested Laplace approximations.
- This doesn't work particularly well without further attention to the marginals.

Finding the latent marginal posterior: [RMC09, Eq. 12]

- Instead focus the Laplace approximation directly on the marginal itself:

$$q_{LA}(f_i|\theta, D) \quad \propto \quad \frac{p(\hat{f}, \theta, D)}{q_{\mathcal{N}}(\hat{f}_{-i}|f_i, \theta, D)},$$

- with the same logic as with $q(\theta|D)$ [RMC09, Eq. 3].
- And thus the latent marginal posterior:

$$\begin{aligned} p(f_i|D) &= \int p(f_i|\theta, D)p(\theta|D)d\theta \\ &\approx \int q_{LA}(f_i|\theta, D)q(\theta|D) \\ &\approx \sum_{k=1}^{K} q_{LA}(f_i|\theta_k, D)q(\theta_k|D), \end{aligned}$$

- Works well but is very computationally inefficient. Why?
- Thereafter some mechanics to create "simplified" Laplace approximation.

hw2: comment on [RMC09]

- [RMC09] was a *read* paper at JRSSB.
- Approximately twenty thorough commentaries are written by world experts.
- Option 1: write your own commentary about the pros and cons of this work.
- Option 2: summarize the opinions of one commentator (or a group of commentators), and then give your own meta-commentary on those opinions.
- Option 2 is probably easier and more rewarding.
- To do: write ~ 1 page, due next Friday 9 October at noon (on courseworks).

References

[KR05] Malte Kuss and Carl Edward Rasmussen.
Assessing approximate inference for binary gaussian process classification.
The Journal of Machine Learning Research, 6:1679–1704, 2005.

[MA10] Iain Murray and Ryan P Adams.
Slice sampling covariance hyperparameters of latent gaussian models.
In *Advances in Neural Information Processing Systems*, pages 1732–1740, 2010.

[MAM09] Iain Murray, Ryan Prescott Adams, and David JC MacKay.
Elliptical slice sampling.
arXiv preprint arXiv:1001.0175, 2009.

[Nea98] Radford M Neal.
Regression and classification using gaussian process priors.
Bayesian statistics, 6, 1998.

[RMC09] Havard Rue, Sara Martino, and Nicolas Chopin.
Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.
Journal of the royal statistical society: Series b (statistical methodology), 71(2):319–392, 2009.

[RW06] C. E. Rasmussen and C.K.I. Williams.
Gaussian Processes for Machine Learning.
MIT Press, Cambridge, 2006.