Goals: summarize posterior in simple form, estimate model evidence $P(D|\mathcal{M})$.

Posterior distributions

$$p(\theta|D, \mathcal{M}) = \frac{P(D|\theta) p(\theta)}{P(D)}$$

E.g., logistic regression:

$$p(\theta) = \mathcal{N}(\theta; 0, \sigma^2 I)$$

$$P(D|\theta) = \prod \sigma(z^{(n)} w^T x^{(n)}), \quad \text{labels } z^{(n)} \in \{\pm 1\}$$

Used to approximate model likelihood (AKA 'evidence', 'marginal likelihood'):

$$P(D) = \int P(D|\theta) p(\theta) d\theta$$

Integrate large product non-linear functions.

Posterior after 500 datapoints

$N=50$ labels generated with $w=1$ at $z^{(n)} \sim \mathcal{N}(0,10^2)$

$$p(w) \propto \mathcal{N}(w; 0,1)$$

$$p(w|D) \propto \mathcal{N}(w; 0,1) \sigma(10-20w)$$

Finite parameter vector $\theta$

$P(\theta|\text{lots of data})$ often nearly Gaussian around the mode

Need to identify which Gaussian it is: mean, covariance

Gaussian approximations

MAP estimate:

$$\theta^* = \arg \max \theta \left[ \log P(D|\theta) + \log P(\theta) \right].$$

Taylor expand at optimum:

$$-\log P(\theta|D) = E(\theta) = -\log P(D|\theta) - \log P(\theta) + \log P(D).$$

Because $\nabla \theta E$ is zero at $\theta^*$ (a turning point):

$$E(\theta^* + \delta) \approx E(\theta^*) + \frac{1}{2} \delta^T H \delta$$

Do same thing to Gaussian around mean, identify Laplace approximation:

$$P(\theta|D) \approx \mathcal{N}(\theta; \theta^*, H^{-1})$$

Deterministic Approximations 2

Laplace and variational approximations

Laplace Approximation

MAP estimate:

Weird densities (we’ve seen sometimes happen) won’t work well. We only locally match one mode.

Mode may not have much mass, or misleading curvature

High dimensions: mode may be flat in some direction

→ill-conditioned Hessian

Laplace details

Matrix of second derivatives is called the Hessian:

$$H_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \left[ -\log P(\theta|D) \right]_{\theta=\theta^*}$$

Find posterior mode (MAP estimate) $\theta^*$ using favourite gradient-based optimizer.

Log posterior doesn’t need to be normalized: constants disappear from derivatives and second-derivatives

Laplace picture

Curvature and mode match.

We can normalize Gaussian. Height at mode won’t match exactly!

Used to approximate model likelihood (AKA ‘evidence’, ‘marginal likelihood’):

$$P(D) = \frac{P(D|\theta) p(\theta)}{P(D)} \approx \frac{P(D|\theta^*) p(\theta^*)}{\mathcal{N}(\theta^*; \theta^*, H^{-1})} = P(D|\theta^*) P(\theta^*) |2\pi H^{-1}|^{\frac{1}{2}}$$

Laplace problems

Weird densities (we’ve seen sometimes happen) won’t work well. We only locally match one mode.

Mode may not have much mass, or misleading curvature

High dimensions: mode may be flat in some direction

→ill-conditioned Hessian
### Variational methods

**Goal:** fit target distribution (e.g., parameter posterior)

**Define:**
- family of possible distributions $q(\theta)$
- ‘variational objective’ (says ‘how well does $q$ match?’)

**Optimize objective:**
- Fit parameters of $q(\theta)$ — e.g., mean and cov of Gaussian
- Capturing posterior width better than only fitting point estimate

**DKL $(p\|q)$:** fitting posterior

Fit $q$ to $p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$

Substitute into KL and get spray of terms:

$$D_{KL}(q||p) = E_q[\log q(\theta)] - E_q[\log p(D|\theta)] - E_q[\log p(\theta)] + \log p(D)$$

First three terms: Minimize sum of these, $J(q)$.

$\log p(D)$: Model evidence. Usually intractable, but:

$$D_{KL}(q||p) \geq 0 \Rightarrow \log p(D) \geq -J(q)$$

We optimize lower bound on the log marginal likelihood

**DKL $(q||p)$:** optimization

Literature full of clever (non-examinable) iterative ways to optimize $D_{KL}(q||p)$. $q$ not always Gaussian.

Use standard optimizers? Hardest term to evaluate is:

$$E_\theta[\log p(D|\theta)] = \sum_{n=1}^{N} E_\theta[\log p(x_n|\theta)]$$

Sum of possibly simple integrals.

Stochastic gradient descent is an option.

---

### Summary

**Laplace approximation:**
- Straightforward to apply; accuracy variable
- 2nd derivatives $\rightarrow$ certainty of parameter
- Incremental improvement on MAP estimate

**Variational methods:**
- Fit variational parameters of $q$ (not $\theta$)
- $KL(p||q)$ vs. $KL(q||p)$
- Bound marginal/model likelihood (‘the evidence’)