STATS271/371: Applied Bayesian Statistics

Lap 2: Generalized Linear Models and the Laplace Approximation

Scott Linderman

April 12, 2021

Announcements

- Lecture slides and course content on Github: `https://github.com/slinderman/stats271sp2021`
- Homework 2 due Friday (April 16).

Final Projects

- **Goal:** Go through Box's Loop on a real dataset.
  - Find an interesting real dataset, e.g. from your own research, a recent paper, or a public repository. (Synthetic datasets are not sufficient.)
  - Develop a model.
  - Think carefully about Bayesian inference algorithms and run one.
  - Evaluate your model, then revise it in light of your analysis.
  - Repeat...
  - Present your results and findings in the form of a short report.
- You may work individually or in a team of two.
- Proposals due late April/early May (TBD).

Survey Results

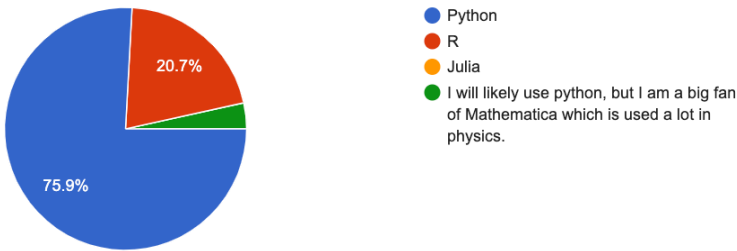
Where are you zooming in from?

- Campus or nearby: 24/29 (!!!)
- San Francisco
- India
- Senatobia, MS
- North Carolina
- Houston, TX

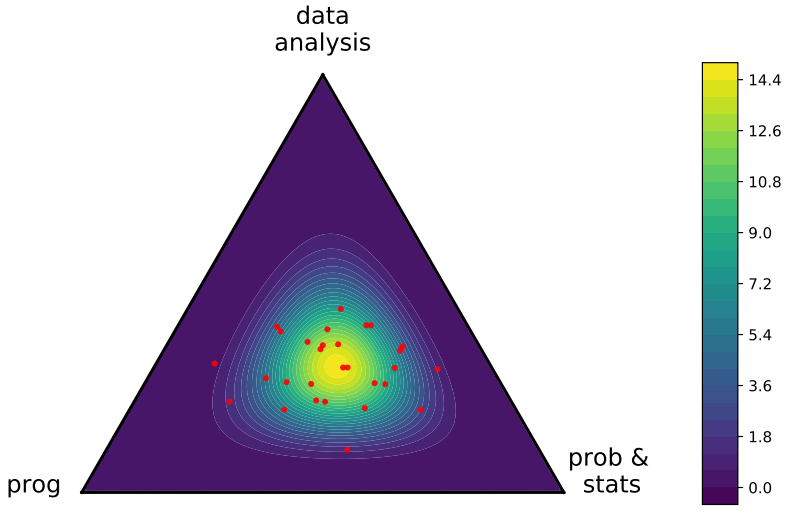
Survey Results II

What's your programming language of choice?

29 responses



Survey Results III

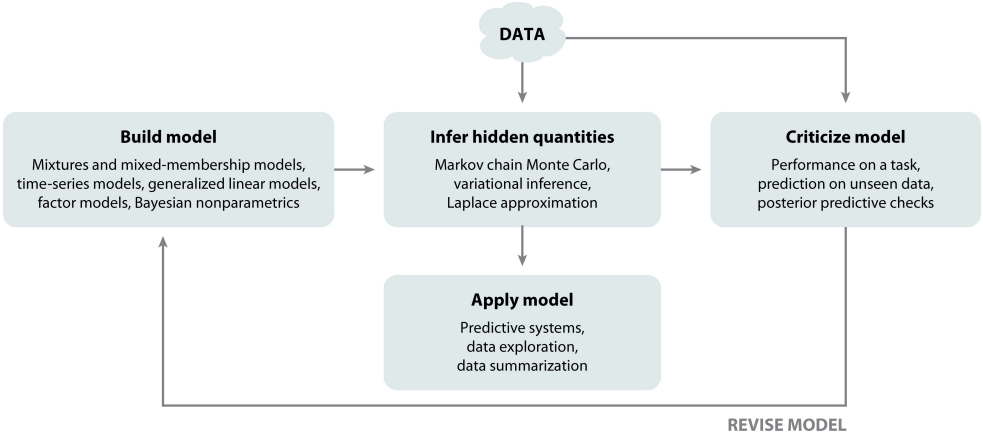


Survey Results IV

Any pandemic hobbies?

- Fingerstyle guitar
- Cycling
- Hiking
- Playing music
- Rock hard abs
- Golf
- Jouranling / Tik Tok
- Tennis (x3)
- Checking Covid numbers :(
- Cooking (x3)
- Reading (x2)
- Looking after my newborn :)
- Astrophotography
- Yoga/Pokemon Go
- Chess
- Learning Spanish
- Dominion
- Walking
- Piano

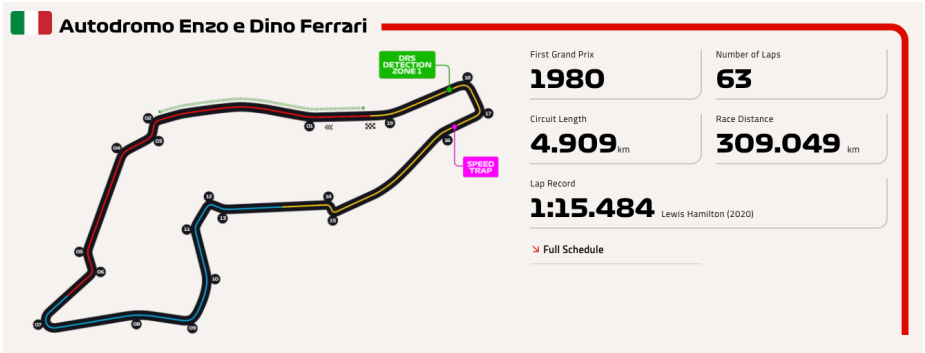
Lap 2 of Box's Loop



Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

Blei, *Ann. Rev. Stat. App.* 2014.

Lap 2 of Box's Loop



https://www.formula1.com/en/racing/2021/EmiliaRomagna/Circuit.html

Bayesian Generalized Linear Models

Lap 2 of Box's loop will introduce:

- **Model:** Bayesian generalized linear models
- **Algorithm:** Laplace approximation
- **Criticism:** Posterior predictive checks

Motivation

Linear regression assumed real-valued observations, but what if our data has some other support?

- $y_n \in \{0, 1\}$ binary observation; e.g. win/lose, healthy/sick.
- $y_n \in \{0, 1, 2, \ldots\}$ count observation; e.g. number of spikes in 25ms bin.
- $y_n \in \{0, 1, \ldots, M\}$ count observation (with max); e.g. number of races won out of M total
- $y_n \in \{1, \ldots, K\}$ unordered, categorical observation; e.g.
  - Democrat/Republican/Independent
  - Hamilton/Madison/Jay (authors of the Federalist papers)
  - Hamilton/Verstappen/Bottas (only F1 drivers who win races)
  - Stanford/Berkeley/Not-Stanford-or-Berkeley (only institutions of higher learning)

We could still apply linear regression, but it would make weird predictions. *There will be -2 spikes in the next time bin? Bet on Hamilstappen to win?*

Notation

Let

- $y_n \in \mathscr{Y}$ denote the *n*-th *observation* (type to be specified later)
- $\boldsymbol{x}_n \in \mathbb{R}^P$ denote a the *covariates* (aka features) correspond the *n*-th datapoint
- $\boldsymbol{w} \in \mathbb{R}^P$ denote the *weights* of the model

Example: Bernoulli GLM

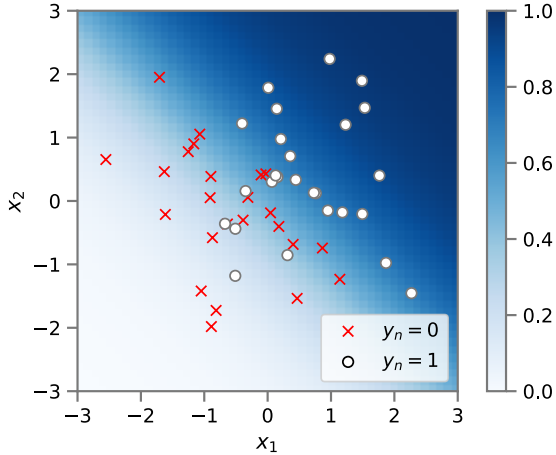
- For example, consider modeling binary observations $y_n \in \{0, 1\}$ given covariates $\boldsymbol{x}_n \in \mathbb{R}^P$.
- We model the observations as Bernoulli random variables,

$$y_n \mid \boldsymbol{x}_n \sim \text{Bern}(p_n) \tag{1}$$

where

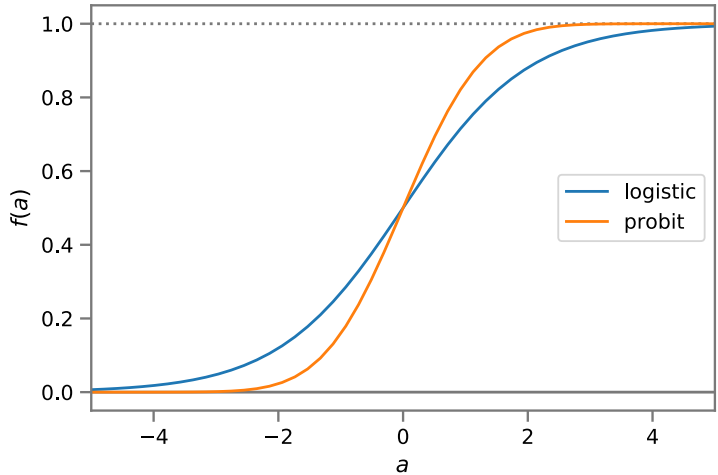
$$p_n = \Pr(y_n = 1 \mid \boldsymbol{x}_n) = \mathbb{E}[y_n \mid \boldsymbol{x}_n]. \tag{2}$$

- Then, model the conditional expectation as $\mathbb{E}[y_n \mid \boldsymbol{x}_n] = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ where $f : \mathbb{R} \to [0, 1]$ is the *mean function*.



Logistic and probit functions

Common choices for f include the logistic and probit functions, corresponding to the *logistic regression* and *probit regression* models, respectively.



Likelihood

The likelihood of N conditionally independent observations is,

$$p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{w}) = \prod_{n=1}^N \text{Bern}(y_n \mid f(\boldsymbol{w}^\top \boldsymbol{x}_n)) \tag{3}$$

$$= \prod_{n=1}^N f(\boldsymbol{w}^\top \boldsymbol{x}_n)^{y_n} \left(1 - f(\boldsymbol{w}^\top \boldsymbol{x}_n)\right)^{1-y_n} \tag{4}$$

For example, consider the *logistic function* $f(a) = \frac{e^a}{1+e^a}$. Then,

$$p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{w}) = \prod_{n=1}^N \left[\frac{e^{\boldsymbol{w}^\top \boldsymbol{x}_n}}{1 + e^{\boldsymbol{w}^\top \boldsymbol{x}_n}}\right]^{y_n} \left[\frac{1}{1 + e^{\boldsymbol{w}^\top \boldsymbol{x}_n}}\right]^{1-y_n} \tag{5}$$

$$= \prod_{n=1}^N \frac{e^{\langle y_n \boldsymbol{x}_n, \boldsymbol{w}\rangle}}{1 + e^{\boldsymbol{w}^\top \boldsymbol{x}_n}} \tag{6}$$

The numerator can be written in terms of $\sum_n y_n \boldsymbol{x}_n$, but the denominator doesn't easily simplify.

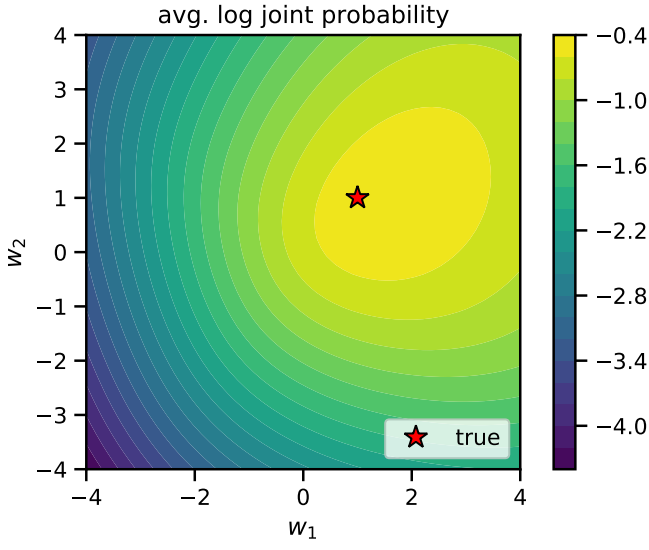
Prior

- Here, we'll use a spherical multivariate normal prior,

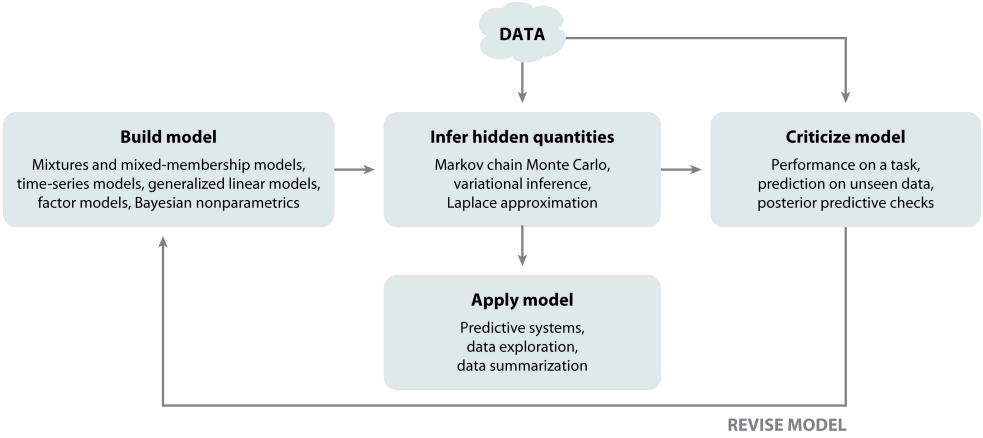
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \sigma^2 \boldsymbol{I}), \tag{7}$$

  where $\mathbf{0} \in \mathbb{R}^P$ and $\sigma^2 \in \mathbb{R}_+$ are the mean and variance, respectively.
- This is the simplest choice, but we have lots of flexibility if a stronger prior is warranted.
- For example, it's easy to generalize to arbitrary mean and covariance, μ and Σ.

Log probability contours



Box's Loop: Infer hidden quantities



Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

Blei, *Ann. Rev. Stat. App.* 2014.

Maximum a posteriori (MAP) estimation

Before going to full Bayesian inference, let's just look for the posterior mode.

The log posterior probability is equal to the log joint minus the log marginal likelihood, which is constant with respect to the parameters $\boldsymbol{w}$,

$$\mathscr{L}(\boldsymbol{w}) = \log p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{w}) + \log p(\boldsymbol{w}) - \underbrace{\log p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N)}_{\text{const. w.r.t. } \boldsymbol{w}} \tag{8}$$

$$= \sum_{n=1}^N \log \frac{e^{\langle y_n \boldsymbol{x}_n, \boldsymbol{w} \rangle}}{1 + e^{\boldsymbol{w}^\top \boldsymbol{x}_n}} - \frac{1}{2\sigma^2} \boldsymbol{w}^\top \boldsymbol{w} + c \tag{9}$$

$$= \left\langle \sum_{n=1}^N y_n \boldsymbol{x}_n, \boldsymbol{w} \right\rangle - \sum_{n=1}^N \log(1 + \exp\{\boldsymbol{x}_n^\top \boldsymbol{w}\}) - \frac{1}{2\sigma^2} \boldsymbol{w}^\top \boldsymbol{w} + c \tag{10}$$

Gradient of the log probability

Taking the gradient and setting it to zero,

$$\nabla_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}) = \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} \frac{\exp\{\boldsymbol{x}_n^\top \boldsymbol{w}\}}{1 + \exp\{\boldsymbol{x}_n^\top \boldsymbol{w}\}} \boldsymbol{x}_n - \frac{1}{\sigma^2} \boldsymbol{w} \tag{11}$$

$$= \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} f(\boldsymbol{x}_n^\top \boldsymbol{w}) \boldsymbol{x}_n. - \frac{1}{\sigma^2} \boldsymbol{w} \tag{12}$$

In matrix notation,

$$\boldsymbol{X} = \begin{bmatrix} - & \boldsymbol{x}_1^\top & - \\ & \vdots & \\ - & \boldsymbol{x}_N^\top & - \end{bmatrix}, \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \tag{13}$$

and letting $\hat{\boldsymbol{y}} = f(\boldsymbol{X}\boldsymbol{w})$, where f is applied elementwise, we can write the gradient as,

$$\nabla_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top \boldsymbol{X} - \frac{1}{\sigma^2} \boldsymbol{w}. \tag{14}$$

Gradient of the log probability II

- In other words, the *gradient of the log likelihood is a weighted sum of the covariates* where the weights are given by the error $y_n - \hat{y}_n$.
- Note that the gradient is still a *nonlinear function of* $\boldsymbol{w}$ since $\hat{y} = f(\boldsymbol{x}_n^\top \boldsymbol{w})$.
- The nonlinear equation $\nabla_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}) = 0$ does not have a closed form solution.

Hessian of the log probability

- The Hessian—i.e. the matrix of second partial derivatives, or the Jacobian of the gradient—is,

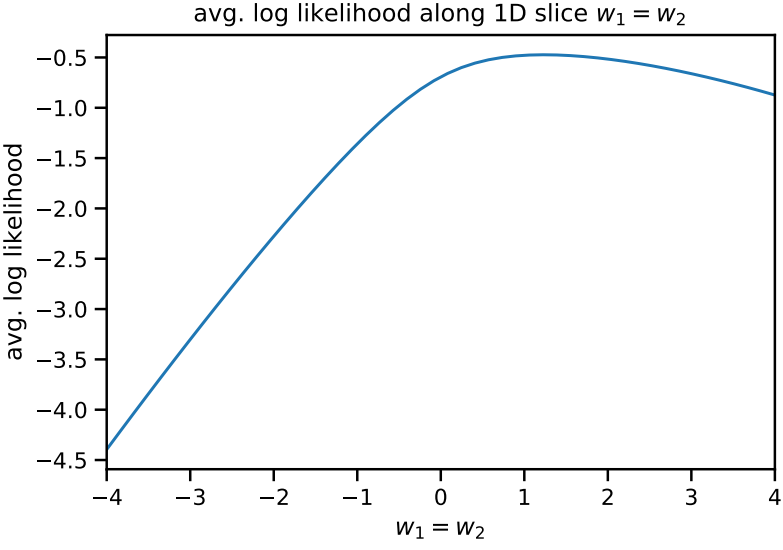
$$\nabla^2_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \sum_{n=1}^{N} \Big(y_n - f(\boldsymbol{x}_n^\top \boldsymbol{w}) \Big) \boldsymbol{x}_n - \frac{1}{\sigma^2} \boldsymbol{w}. \tag{15}$$

$$= -\left(\sum_{n=1}^{N} f'(\boldsymbol{x}_n^\top \boldsymbol{w}) \boldsymbol{x}_n \boldsymbol{x}_n^\top + \frac{1}{\sigma^2} \boldsymbol{I} \right) \tag{16}$$

  where $f' : \mathbb{R} \to \mathbb{R}$ is the derivative of f.

- Since f is monotonically increasing, $f' \geq 0$.
- The negative Hessian is a weighted sum of outer products with positive weights $f'(\boldsymbol{x}_n^\top \boldsymbol{w})$. Thus, the Hessian is negative definite.
- That means the *log probability is concave* and has a global optimum.

Concavity of the log probability



Exercise: The derivative of the logistic function

Show that

$$f'(a) = \left[\frac{\mathrm{d}}{\mathrm{d}\alpha}\frac{e^{\alpha}}{1+e^{\alpha}}\right]_{\alpha=a} = f(a)(1-f(a)) = f(a)f(-a) \geq 0. \tag{17}$$

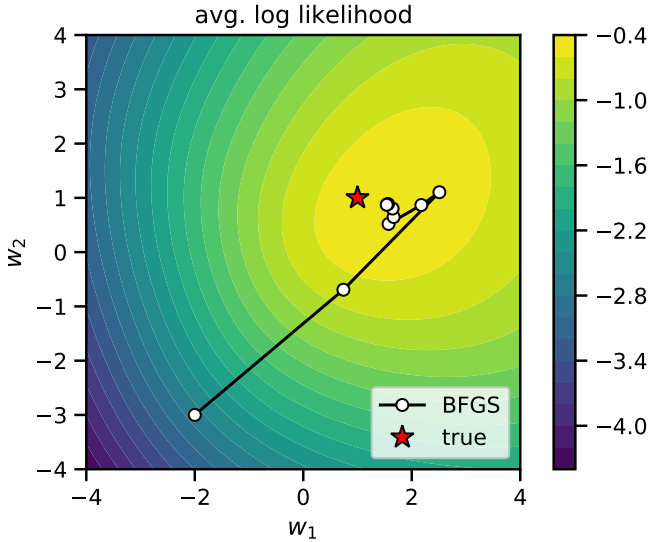
Algorithm

- Though the MAP estimate doesn't have a closed form solution, we can find it via optimization.
- Since the log probability is concave, standard optimization methods will find the global optimum.
- For example, (damped) *Newton's method*,

$$\boldsymbol{w}^{(i+1)} \leftarrow \boldsymbol{w}^{(i)} - \gamma[\nabla^2_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}^{(i)})]^{-1} \nabla_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}^{(i)}) \tag{18}$$

  (with sufficiently small step size $\gamma \in (0, 1]$) achieves second-order convergence to the optimum, $\boldsymbol{w}_{\text{MAP}}$.
- With many covariates (when P is large), Newton's method incurs an $O(P^3)$ complexity.
- *Quasi*-Newton methods, like the BFGS algorithm, build an approximation to approximate Newton's method at lower computational cost.

BFGS Convergence



Posterior Distribution

What if we want more than a point estimate?

The posterior distribution is given by,

$$p(\boldsymbol{w} \mid \{\boldsymbol{x}_n, y_n\}_{n=1}^N) = \frac{p(\boldsymbol{w})\, p(\{y_n\}_{n=1}^N \mid \boldsymbol{w}, \{\boldsymbol{x}_n\}_{n=1}^N)}{p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N)} \tag{19}$$

$$\tag{20}$$

Unfortunately, the marginal likelihood is hard to compute since

$$p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N) = \int \prod_{n=1}^N \mathrm{Bern}(y_n \mid f(\boldsymbol{w}^\top \boldsymbol{x}_n))\, \mathcal{N}(\boldsymbol{w} \mid 0, \sigma^2 \boldsymbol{I})\, \mathrm{d}\boldsymbol{w} \tag{21}$$

does not have a closed form solution.

Laplace Approximation

Idea: *approximate the posterior with a multivariate normal distribution centered on the mode.*

To motivate this, consider a second-order Taylor approximation to the log posterior,

$$\mathscr{L}(\boldsymbol{w}) \approx \mathscr{L}(\boldsymbol{w}_{\text{MAP}}) + (\boldsymbol{w} - \boldsymbol{w}_{\text{MAP}})^\top \underbrace{\nabla_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}_{\text{MAP}})}_{\mathbf{0} \text{ at the mode}} + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\text{MAP}})^\top \nabla^2_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}_{\text{MAP}})(\boldsymbol{w} - \boldsymbol{w}_{\text{MAP}}) \tag{22}$$

$$= -\frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\text{MAP}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w} - \boldsymbol{w}_{\text{MAP}}) + c \tag{23}$$

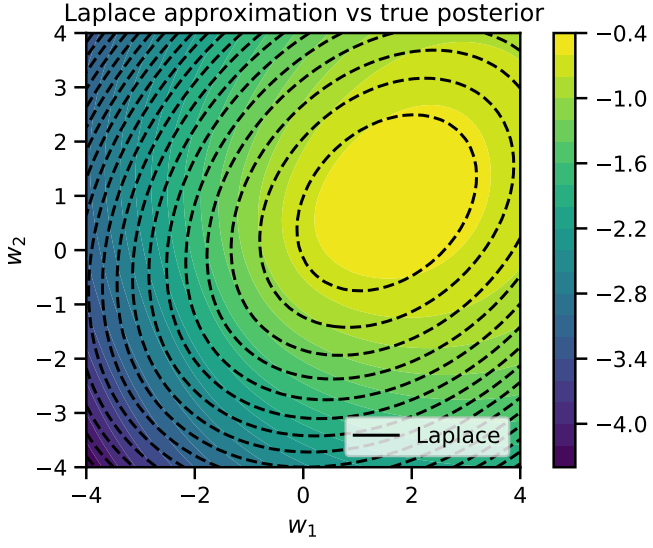
$$= \log \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{w}_{\text{MAP}}, \boldsymbol{\Sigma}) \tag{24}$$

where $\boldsymbol{\Sigma} = -[\nabla^2_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}_{\text{MAP}})]^{-1}$

In other words, the posterior is approximately Gaussian with covariance given by the (negative) inverse Hessian at the mode.

Since the Hessian is *negative* definite, the covariance is *positive* definite, as required.

Laplace Approximation II



Bernstein-von Mises Theorem

In the large data limit (as $N \to \infty$), the posterior is asymptotically normal, justifying the Laplace approximation in this regime.

Consider a simpler setting in which we have data $\{y_n\}_{n=1}^N \overset{\text{iid}}{\sim} p(y \mid \theta_{\text{true}})$.

Under some conditions (e.g. θ_{true} not on the boundary of Θ and θ_{true} has nonzero prior probability), then the MAP estimate is consistent. As $N \to \infty$, $\theta_{\text{MAP}} \to \theta_{\text{true}}$.

Likewise,

$$p(\theta \mid \{y_n\}_{n=1}^N) \to \mathcal{N}\left(\theta \mid \theta_{\text{true}}, \tfrac{1}{N}[J(\theta_{\text{true}})]^{-1}\right) \tag{25}$$

where

$$J(\theta) = -\mathbb{E}_{p(y|\theta)}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log p(y \mid \theta)\right] \tag{26}$$

is the *Fisher information* of parameter θ.

Approximating the marginal likelihood

The Laplace approximation also offers an approximation of the intractable marginal likelihood,

$$\begin{aligned}\mathscr{L}(\boldsymbol{w}_{\text{MAP}}) &= \log p(\boldsymbol{w}_{\text{MAP}}) + \log p(\{y_n\}_{n=1}^N \mid \boldsymbol{w}_{\text{MAP}}, \{\boldsymbol{x}_n\}_{n=1}^N) - \log p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N) && (27)\\ &\approx \log \mathcal{N}(\boldsymbol{w}_{\text{MAP}} \mid \boldsymbol{w}_{\text{MAP}}, \boldsymbol{\Sigma}) && (28)\\ &= -\frac{P}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| && (29)\end{aligned}$$

Rearranging terms,

$$\begin{aligned}\log p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N) \approx \log p(\boldsymbol{w}_{\text{MAP}}) + \log p(\{y_n\}_{n=1}^N \mid \boldsymbol{w}_{\text{MAP}}, \{\boldsymbol{x}_n\}_{n=1}^N)\\ + \frac{P}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| \qquad (30)\end{aligned}$$

Note: combine this with $\boldsymbol{\Sigma} \approx \frac{1}{N}[J(\boldsymbol{w}_{MAP})]^{-1}$ and $\frac{1}{2}\log|\boldsymbol{\Sigma}| \approx \frac{P}{2}\log N + O(1)$ to derive the *Bayesian information criterion (BIC)*, a technique for penalized maximum likelihood estimation.

Exponential family distributions

Now we'll consider GLMs with *exponential family* observations,

$$p(y_n \mid \eta_n) = h(y_n) \exp\left\{\langle t(y_n), \eta_n \rangle - A(\eta_n)\right\}, \tag{31}$$

where

- $h(y_n) : \mathcal{Y} \to \mathbb{R}_+$ is the *base measure*,
- $t(y_n) \in \mathbb{R}^T$ are the *sufficient statistics*,
- $\eta_n \in \mathbb{R}^T$ are the *natural parameters*, and
- $A(\eta_n) : \mathbb{R}^T \to \mathbb{R}$ is the *log normalizing* function (aka the *partition function*).

Example: Bernoulli distribution in exponential family form

The Bernoulli distribution is a special case,

$$\begin{aligned}
\text{Bern}(y_n \mid p_n) &= p_n^{y_n}(1-p_n)^{1-y_n} && (32)\\
&= \exp\left\{y_n \log p_n + (1-y_n)\log(1-p_n)\right\} && (33)\\
&= \exp\left\{y_n \log \frac{p_n}{1-p_n} + \log(1-p_n)\right\} && (34)\\
&= h(y_n)\exp\left\{y_n \eta_n - A(\eta_n)\right\} && (35)
\end{aligned}$$

where

$$h(y_n) = 1 \qquad t(y_n) = y_n \qquad \eta_n = \log\frac{p_n}{1-p_n} \qquad A(\eta_n) = -\log(1-p_n) \quad (36)$$

$$= -\log\left(1 - \frac{e^{\eta_n}}{1+e^{\eta_n}}\right) \quad (37)$$

$$= \log\left(1+e^{\eta_n}\right). \quad (38)$$

Gradient of the log normalizer yields the expected sufficient statistics

By definition,

$$A(\eta_n) = \log \int h(y_n) \exp\left\{\langle t(y_n), \eta_n \rangle\right\} \mathrm{d}y_n \tag{39}$$

so

$$\nabla_{\eta_n} A(\eta_n) = \nabla_{\eta_n} \log \int h(y_n) \exp\left\{\langle t(y_n), \eta_n \rangle\right\} \mathrm{d}y_n \tag{40}$$

$$= \frac{\int h(y_n) \exp\left\{\langle t(y_n), \eta_n \rangle\right\} t(y_n) \, \mathrm{d}y_n}{\int h(y_n) \exp\left\{\langle t(y_n), \eta_n \rangle\right\} \mathrm{d}y_n} \tag{41}$$

$$= \int p(y_n \mid \eta_n) \, t(y_n) \, \mathrm{d}y_n \tag{42}$$

$$= \mathbb{E}_{p(y_n|\eta_n)}[t(y_n)] \tag{43}$$

Hessian of the log normalizer yields the covariance of the sufficient statistics

Likewise,

$$\nabla^2_{\eta_n} A(\eta_n) = \nabla_{\eta_n} \int p(y_n \mid \eta_n)\, t(y_n)\, \mathrm{d}y_n \tag{44}$$

$$= \int p(y_n \mid \eta_n)\, t(y_n)\, (t(y_n) - \nabla_{\eta_n} A(\eta_n))^\top \, \mathrm{d}y_n \tag{45}$$

$$= \mathbb{E}_{p(y_n|\eta_n)}[t(y_n) t(y_n)^\top] - \mathbb{E}_{p(y_n|\eta_n)}[t(y_n)]\, \mathbb{E}_{p(y_n|\eta_n)}[t(y_n)]^\top \tag{46}$$

$$= \mathrm{Cov}_{p(y_n|\eta_n)}[t(y_n)] \tag{47}$$

Covariances are positive semi-definite, so the *log normalizer is a convex function*.

Exponential family GLMs

- To construct a generalized linear model with exponential family observations, we set

$$\mathbb{E}[y_n \mid \boldsymbol{x}_n] = f(\boldsymbol{w}^\top \boldsymbol{x}_n). \tag{48}$$

- From above, this implies,

$$\nabla_{\eta_n} A(\eta_n) = f(\boldsymbol{w}^\top \boldsymbol{x}_n) \tag{49}$$

$$\Rightarrow \eta_n = [\nabla A]^{-1}\big(f(\boldsymbol{w}^\top \boldsymbol{x}_n)\big), \tag{50}$$

  when $\nabla A(\cdot)$ is invertible. (In this case, the exponential family is said to be *minimal*).

- The *canonical mean function* is $f(\cdot) = \nabla A(\cdot)$ so that $\eta_n = \boldsymbol{w}^\top \boldsymbol{x}_n$.
- The (canonical) *link function* is the inverse of the (canonical) mean function.

Logistic regression revisited

Consider the Bernoulli distribution once more. The gradient of the log normalizer is,

$$\nabla_\eta A(\eta) = \nabla_\eta \log(1 + e^\eta) = \frac{e^\eta}{1 + e^\eta} \tag{51}$$

This is the logistic function!

Thus, logistic regression with is a Bernoulli GLM with the canonical mean function.

Why care about canonical mean functions?

(**Spoiler:** It's 2021 and we have automatic differentiation so it's not such a big deal anymore.)

Canonical mean functions lead to nice math. Consider the log joint probability,

$$\mathscr{L}(\boldsymbol{w}) = \log p(\boldsymbol{w}) + \sum_{n=1}^{N} \langle t(y_n), \eta_n \rangle - A(\eta_n) + c \tag{52}$$

$$= -\frac{1}{2\sigma^2} \boldsymbol{w}^\top \boldsymbol{w} + \sum_{n=1}^{N} \langle t(y_n), \boldsymbol{w}^\top \boldsymbol{x}_n \rangle - A(\boldsymbol{w}^\top \boldsymbol{x}_n) + c, \tag{53}$$

where we have assumed $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and canonical mean function so $\eta_n = \boldsymbol{w}^\top \boldsymbol{x}_n$.

The gradient is,

$$\nabla_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}) = -\frac{1}{\sigma^2} \boldsymbol{w} + \sum_{n=1}^{N} \langle t(y_n), \boldsymbol{x}_n \rangle - \langle \nabla A(\boldsymbol{w}^\top \boldsymbol{x}_n), \boldsymbol{x}_n \rangle \tag{54}$$

$$= -\frac{1}{\sigma^2} \boldsymbol{w} + \sum_{n=1}^{N} \langle t(y_n) - \mathbb{E}_{p(y_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n)}[t(y_n)], \boldsymbol{x}_n \rangle \tag{55}$$

Why care about canonical mean functions? II

In many cases, $t(y_n) = y_n \in \mathbb{R}$ so

$$\nabla_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}) = -\frac{1}{\sigma^2}\boldsymbol{w} + \sum_{n=1}^{N}(y_n - \hat{y}_n)\boldsymbol{x}_n. \tag{56}$$

And in that case the Hessian is

$$\nabla^2_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}) = -\frac{1}{\sigma^2}\boldsymbol{I} - \sum_{n=1}^{N} \nabla^2 A(\boldsymbol{w}^\top \boldsymbol{x}_n)\, \boldsymbol{x}_n \boldsymbol{x}_n^\top \tag{57}$$

$$= -\left(\frac{1}{\sigma^2}\boldsymbol{I} + \sum_{n=1}^{N} \mathrm{Var}_{p(y_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n)}[y_n]\, \boldsymbol{x}_n \boldsymbol{x}_n^\top \right) \tag{58}$$

Why care about canonical mean functions? III

Now recall the Newton's method updates, written here in terms of the change in weights,

$$\Delta \boldsymbol{w} = -[\nabla^2_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w})]^{-1} \nabla_{\boldsymbol{w}} \mathscr{L}(\boldsymbol{w}) \tag{59}$$

$$= \left[\frac{1}{\sigma^2} \boldsymbol{I} + \sum_{n=1}^{N} \mathrm{Var}_{p(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n)}[y_n] \, \boldsymbol{x}_n \boldsymbol{x}_n^\top \right]^{-1} \left[-\frac{1}{\sigma^2} \boldsymbol{w} + \sum_{n=1}^{N} (y_n - \hat{y}_n) \boldsymbol{x}_n \right] \tag{60}$$

Letting $\hat{v}_n = \mathrm{Var}_{p(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n)}[y_n]$ and taking the uninformative limit of $\sigma^2 \to \infty$,

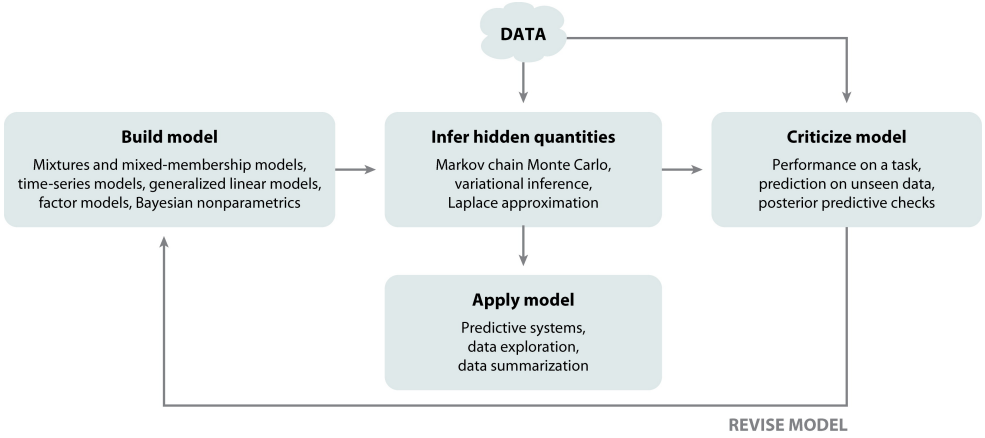
$$\Delta \boldsymbol{w} = \left[\sum_{n=1}^{N} \hat{v}_n \boldsymbol{x}_n \boldsymbol{x}_n^\top \right]^{-1} \left[\sum_{n=1}^{N} (y_n - \hat{y}_n) \boldsymbol{x}_n \right] \tag{61}$$

$$= (\boldsymbol{X}^\top \hat{\boldsymbol{V}} \boldsymbol{X})^{-1} [\boldsymbol{X}^\top \hat{\boldsymbol{V}} \hat{\boldsymbol{z}}] \tag{62}$$

where $\hat{\boldsymbol{V}} = \mathrm{diag}([\hat{v}_1, \ldots, \hat{v}_N])$ and $\hat{\boldsymbol{z}} = \hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{y}})$.

This is *iteratively reweighted least squares* (IRLS) with weights $\hat{v}_n$ and targets $\hat{z}_n = \frac{y_n - \hat{y}_n}{\hat{v}_n}$, both of which are functions of the current weights $\boldsymbol{w}$.

Model evaluation via predictive likelihood



Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

Blei, *Ann. Rev. Stat. App.* 2014.

Measures of predictive accuracy

One way to evaluate a model is through the accuracy of its predictions.

For example, given a *point estimate* $\hat{\theta}$, the *mean squared error* summarizes predictive accuracy as,

$$\frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} (y_n - \mathbb{E}[y_n \mid \hat{\theta}])^2 \tag{63}$$

In *probabilistic prediction* (aka *probabilistic forecasting*), we report a distribution over parameters θ. In Bayesian data analysis, this is the posterior distribution $p(\theta \mid \{y_n\}_{n=1}^N)$.

The *posterior predictive distribution* averages over the posterior to make a prediction,

$$p(y_{n'} \mid \{y_n\}_{n=1}^N) = \int p(y_{n'} \mid \theta)\, p(\theta \mid \{y_n\}_{n=1}^N)\, \mathrm{d}\theta, \tag{64}$$

weighting parameters by their posterior probability.

Scoring rules

A *scoring rule* measures the accuracy of a probabilistic prediction.

The *logarithmic scoring rule* is unique in that it is both *proper* and *local*.

The *log predictive likelihood* of new datapoint $y_{n'}$ is,

$$\log p(y_{n'} \mid \{y_n\}_{n=1}^N) = \log \int p(y_{n'} \mid \theta)\, p(\theta \mid \{y_n\}_{n=1}^N)\, \mathrm{d}\theta. \tag{65}$$

Averaging over the distribution of future data

Ideally, we would measure the expected out-of-sample predictive performance,

$$\mathbb{E}_{f(y)}[\log p(y \mid \{y_n\}_{n=1}^N)] = \int \log p(y \mid \{y_n\}_{n=1}^N) f(y) \, \mathrm{d}y \tag{66}$$

where $f(y)$ is the true, but unknown data-generating distribution.

BDA3 calls this the *expected out-of-sample log predictive density* (elpd).

But we don't know f! Instead, we approximate it with,

$$\frac{1}{N_{\text{test}}} \sum_{n'=1}^{N_{\text{test}}} \log p(y_{n'} \mid \{y_n\}_{n=1}^N) \tag{67}$$

for some *test dataset* $\{y_{n'}\}_{n'=1}^{N_{\text{test}}}$.

Estimating out-of-sample predictive accuracy with available data

- Naively, the *within-sample* predictive accuracy takes the test set to be the training set. As such, it over-estimates the the predictive accuracy.
- *Information criteria* like the AIC, DIC, and WAIC try to correct for this bias by penalizing the within-sample estimate.
- *Cross-validation* splits the data into train and test to approximate the predictive likelihood.
- Hardcore Bayesians don't like to withhold data when computing the posterior. Nevertheless, cross-validated predictive log likelihood is the current standard.

Akaike Information Criterion (AIC)

The *Akaike information criterion* (AIC) is based on an asymptotic normal approximation to the posterior.

It corrects for the bias of the within-sample log likelihood by accounting for the number of parameters (an approximation to the amnount of overfitting),

$$-\frac{1}{2}\text{AIC} = \sum_{n=1}^{N} \log p(y_n \mid \theta_{\text{MLE}}) - K, \tag{68}$$

where K is the number of parameters.

The AIC makes some sense for linear models with uniform priors, but not so much for more complicated models.

Deviance Information Criterion (DIC)

The *deviance information criterion* (DIC) makes two changes,

1. Replace the MLE with the posterior mean, $\bar{\theta} = \mathbb{E}_{p(\theta|y)}[\theta]$.
2. Replace K with a bias correction based on the data.

Then,

$$-\frac{1}{2}\text{DIC} = \sum_{n=1}^{N} \log p(y_n \mid \bar{\theta}) - K_{\text{DIC}} \tag{69}$$

where

$$K_{\text{DIC}} = 2\left(\log \sum_{n=1}^{N} p(y_n \mid \bar{\theta}) - \mathbb{E}_{p(\theta|y)}\left[\log \sum_{n=1}^{N} p(y_n \mid \theta)\right]\right) \tag{70}$$

An alternative definition uses,

$$K_{\text{DIC}} = 2\text{Var}_{p(\theta|y)}\left[\log \sum_{n=1}^{N} p(y_n \mid \bar{\theta}\right] \tag{71}$$

Watanabe-Akaike Information Criterion (WAIC)

The *Watanabe-Akaike information criterion* (WAIC) is similar to the DIC with,

$$K_{\text{WAIC}} = 2\left(\log \sum_{n=1}^{N} \mathbb{E}_{p(\theta|y)}[p(y_n \mid \theta)] - \mathbb{E}_{p(\theta|y)}\left[\log \sum_{n=1}^{N} p(y_n \mid \theta)\right]\right) \tag{72}$$

An alternative definition uses,

$$K_{\text{WAIC}} = \sum_{n=1}^{N} \text{Var}_{p(\theta|y)}\left[\log p(y_n \mid \bar{\theta}\right], \tag{73}$$

Then, the WAIC is,

$$-\frac{1}{2}\text{WAIC} = \sum_{n=1}^{N} \log \int p(y_n \mid \theta)\, p(\theta \mid \{y_n\}_{n=1}^{N}\, \mathrm{d}\theta - K_{\text{WAIC}} \tag{74}$$

Leave-one-out cross validation

Idea: approximate out-of-sample predictive accuracy by randomly splitting the training data.

Leave-one-out cross validation (LOOCV) withholds one datapoint out at a time, computes the posterior given the other $N - 1$, and then evaluates predictive likelihood on the held out datapoint.

$$\mathbb{E}_{f(y)}[\log p(y \mid \{y_n\}_{n=1}^N)] \approx \frac{1}{N} \sum_{n=1}^{N} \log p(y_n \mid \{y_m\}_{m \neq n}). \tag{75}$$

For small N (or when using a small number of folds), a bias-correction can be used. (See pg. 175-176 of the book.)

Cross-validated predictive log likelihood estimates are similar to the *jackknife* resampling method in the sense that it is estimating an expectation wrt the unknown data-generating distribution f by resampling the given data.

Thus, it measures the frequentist properties of a Bayesian procedure.