Lecture 11: Gradient Descent and Conjugate Priors

STAT 545: Intro. to Computational Statistics

Vinayak Rao

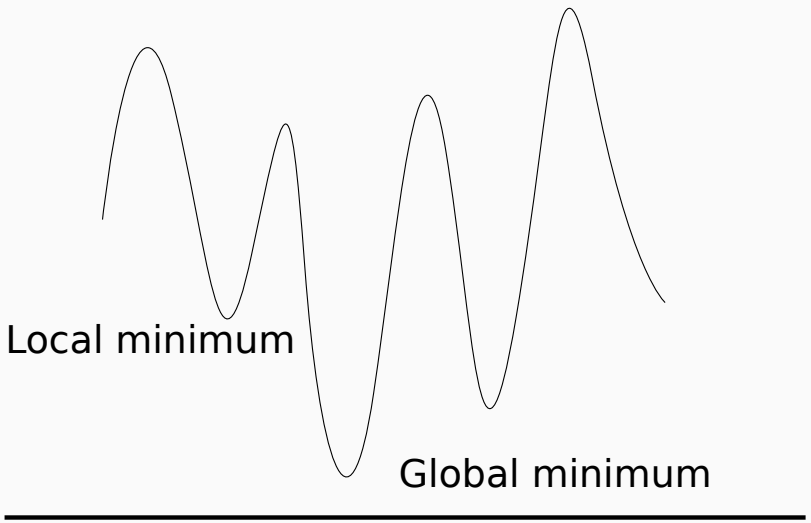
Purdue University

October 6, 2016

Global and local minimum

Find minimum of some function $f : \mathbb{R}^D \to \mathbb{R}$.
(maximization is just minimizing $-f$).

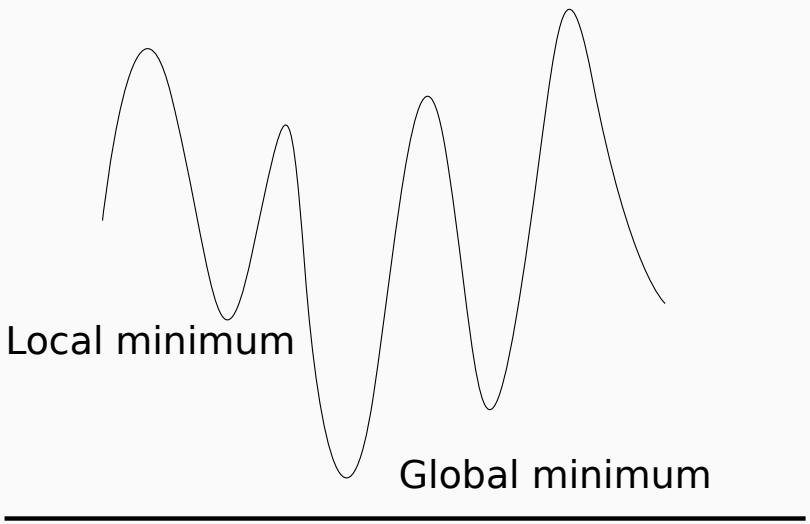
No global information (e.g. only function evaluations, derivatives).



Global and local minimum

Find minimum of some function $f : \mathbb{R}^D \to \mathbb{R}$.
(maximization is just minimizing $-f$).

No global information (e.g. only function evaluations, derivatives).



Finding a global minimum is hard!
Usually settle for finding a local minimum (like the EM algorithm).

Steepest descent (iterative method)

Let x_{old} be our current value.

Update x_{new} as $$x_{new} = x_{old} - \eta \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x_{old}}$$

The steeper the slope, the bigger the move.

Steepest descent (iterative method)

Let x_{old} be our current value.

Update x_{new} as $$x_{new} = x_{old} - \eta \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x_{old}}$$

The steeper the slope, the bigger the move.

η: sometimes called the 'learning rate'
(from neural network literature)

Steepest descent (iterative method)

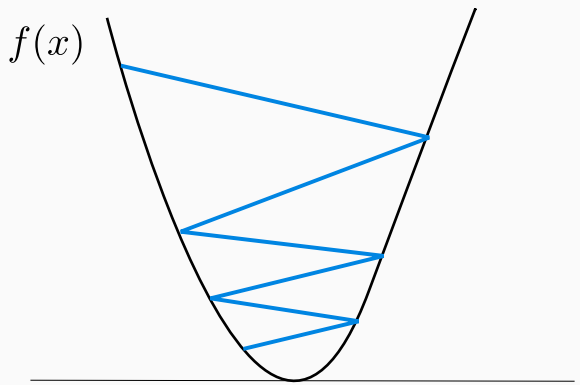
Let x_{old} be our current value.

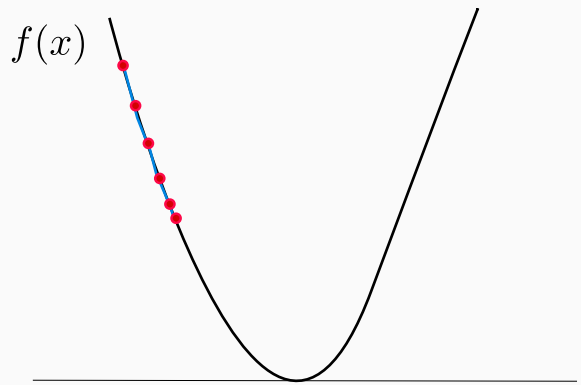
Update x_{new} as $$x_{new} = x_{old} - \eta \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x_{old}}$$

The steeper the slope, the bigger the move.

η: sometimes called the 'learning rate'
(from neural network literature)

Choosing η is a dark art:





Steepest descent (iterative method)

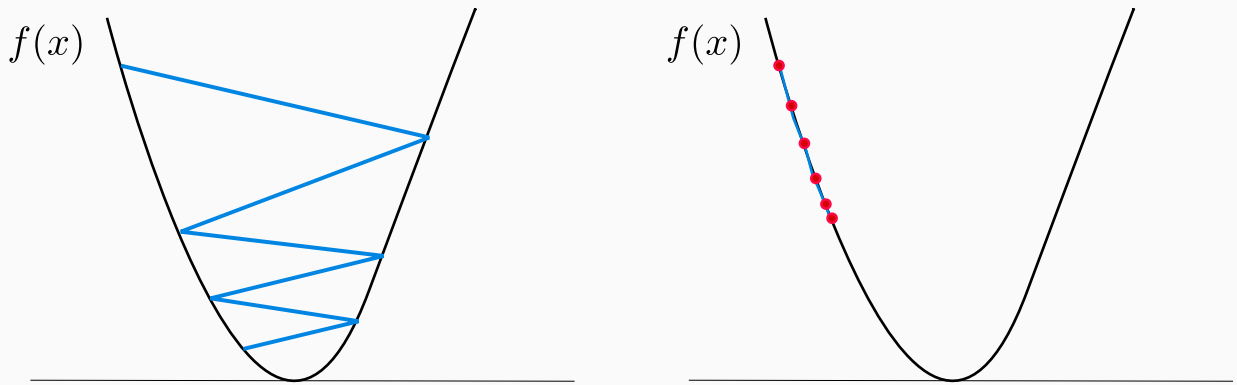
Let x_{old} be our current value.

Update x_{new} as $\qquad x_{new} = x_{old} - \eta \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x_{old}}$

The steeper the slope, the bigger the move.

η: sometimes called the 'learning rate'
(from neural network literature)

Choosing η is a dark art:



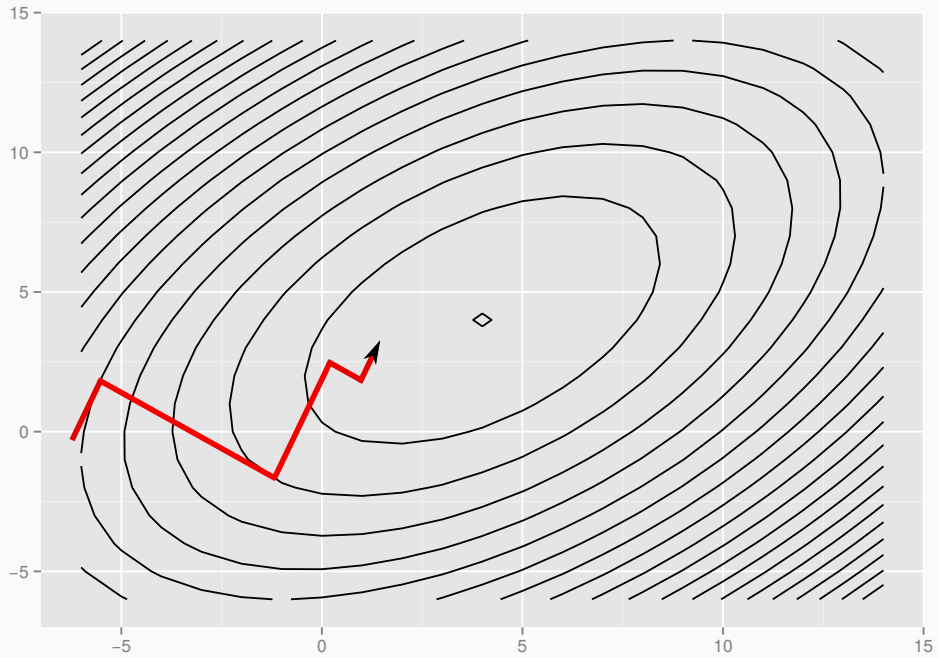
Better methods adapt step-size according to the curvature of f.

Steepest descent in higher-dimensions

Steepest descent also applies to higher dimensions too:

$$x_{new} = x_{old} - \eta \, \nabla f|_{x_{old}}$$

Now, even using the optimal η can be inefficient:



More on this later.

Estimating MLE

Consider a set of observations $Y = (y_1, \cdots, y_N)$.

Assume $y_i \sim p(y|\theta)$

$$\theta_{MLE} = \text{argmax}\ \ell(\theta) = \text{argmax} \sum_{i=1}^{N} \log p(x_i|\theta)$$

Estimating MLE

Consider a set of observations $Y = (y_1, \cdots, y_N)$.

Assume $y_i \sim p(y|\theta)$

$$\theta_{MLE} = \text{argmax}\, \ell(\theta) = \text{argmax} \sum_{i=1}^{N} \log p(x_i|\theta)$$

The gradient of the log-likelihood is $\nabla \ell(\theta) = \sum_{i=1}^{N} \nabla \log p(x_i|\theta)$
The average of the gradients of each datapoint.

Estimating MLE

Consider a set of observations $Y = (y_1, \cdots, y_N)$.

Assume $y_i \sim p(y|\theta)$

$$\theta_{MLE} = \text{argmax}\, \ell(\theta) = \text{argmax} \sum_{i=1}^{N} \log p(x_i|\theta)$$

The gradient of the log-likelihood is $\nabla \ell(\theta) = \sum_{i=1}^{N} \nabla \log p(x_i|\theta)$
The average of the gradients of each datapoint.

Starting with an initial θ_0, iterate:

$$\theta_{i+1} = \theta_i + \eta \nabla \ell(\theta_i)$$

Conceptually (deceptively?) simpler than EM.

Gradient descent (contd.)

$$\nabla \ell(\theta) = \sum_{i=1}^{N} \nabla \log p(x_i|\theta)$$

Cons:

- Calculating the gradient is $O(N)$.
  (Each iteration must cycle through all datapoints.)
- *Lots* of redundancy, esp. for large N.

Gradient descent (contd.)

$$\nabla \ell(\theta) = \sum_{i=1}^{N} \nabla \log p(x_i|\theta)$$

Cons:

- Calculating the gradient is $O(N)$.
  (Each iteration must cycle through all datapoints.)
- *Lots* of redundancy, esp. for large N.

Pros:

- Convergence is better understood.
- Accelerated methods are available (e.g. Newton's method, conjugate gradient)

Stochastic gradient descent

Use a noisy gradient $\widehat{\nabla \ell}$.

Typically split data into N/B batches of size B.
Each iteration, calculate gradient on one of the batches B_i:

$$\widehat{\nabla \ell}(\theta) = \sum_{j \in B_i} \nabla \log p(x_j|\theta)$$

Stochastic gradient descent

Use a noisy gradient $\widehat{\nabla \ell}$.

Typically split data into N/B batches of size B.
Each iteration, calculate gradient on one of the batches B_i:

$$\widehat{\nabla \ell}(\theta) = \sum_{j \in B_i} \nabla \log p(x_j|\theta)$$

Pros:

- Calculating the gradient is $O(B)$.
  (Often, each batch is just a single datapoint)
- Much faster convergence (just one sweep through the data can get you a decent solution).
- Often, you get better solutions.
- Useful for online systems, tracking θ that varies over time .

Stochastic gradient descent

Use a noisy gradient $\widehat{\nabla \ell}$.

Typically split data into N/B batches of size B.
Each iteration, calculate gradient on one of the batches B_i:

$$\widehat{\nabla \ell}(\theta) = \sum_{j \in B_i} \nabla \log p(x_j|\theta)$$

Cons:

- Convergence analysis is harder.
- Noisy gradients mean the algorithm will never converge.
  Typically need to reduce the step size every iteration.
  We want

$$\eta_i \to 0, \quad \sum_{i=1}^{\infty} \eta_i = \infty$$

  E.g. $\eta_i = \frac{a}{b+i}$

BAYESIAN INFERENCE

Bayesian inference

Given a set of observations X, MLE maximizes the likelihood:

$$\theta_{MLE} = \text{argmax } p(X|\theta)$$

Bayesian inference

Given a set of observations X, MLE maximizes the likelihood:

$$\theta_{MLE} = \text{argmax } p(X|\theta)$$

What if we believe θ is close to 0, is sparse, or is smooth?

Encode this with a 'prior' probability $p(\theta)$.

Bayesian inference

Given a set of observations X, MLE maximizes the likelihood:

$$\theta_{MLE} = \text{argmax } p(X|\theta)$$

What if we believe θ is close to 0, is sparse, or is smooth?

Encode this with a 'prior' probability $p(\theta)$.

Given observations, we can calculate the 'posterior':

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{P(X)}$$

Bayesian inference

Given a set of observations X, MLE maximizes the likelihood:

$$\theta_{MLE} = \text{argmax } p(X|\theta)$$

What if we believe θ is close to 0, is sparse, or is smooth?

Encode this with a 'prior' probability $p(\theta)$.

Given observations, we can calculate the 'posterior':

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{P(X)}$$

We can calculate the *maximum a posteriori* (MAP) solution:

$$\theta_{MAP} = \text{argmax } p(\theta|X)$$

Bayesian inference

Given a set of observations X, MLE maximizes the likelihood:

$$\theta_{MLE} = \text{argmax } p(X|\theta)$$

What if we believe θ is close to 0, is sparse, or is smooth?

Encode this with a 'prior' probability $p(\theta)$.

Given observations, we can calculate the 'posterior':

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{P(X)}$$

We can calculate the *maximum a posteriori* (MAP) solution:

$$\theta_{MAP} = \text{argmax } p(\theta|X)$$

Point estimate discards information about uncertainty in θ

Bayesian inference

Bayesian inference works with the entire distribution $p(\theta|X)$.

Allows us to maintain and propagate uncertainty.

Bayesian inference

Bayesian inference works with the entire distribution $p(\theta|X)$.

Allows us to maintain and propagate uncertainty.

In practice, these distributions are unwieldy.

Need approximations.

Bayesian inference

Bayesian inference works with the entire distribution $p(\theta|X)$.

Allows us to maintain and propagate uncertainty.

In practice, these distributions are unwieldy.

Need approximations.

An exception: ‘Conjugate priors’.

Conjugate exponential family priors

Let observations come from an exponential-family:

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta^\top \phi(x))$$
$$= h(x) \exp(\theta^\top \phi(x) - \zeta(\theta)) \quad \text{with } \zeta(\theta) = \log(Z(\theta))$$

Place a prior over θ:

Conjugate exponential family priors

Let observations come from an exponential-family:

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta^\top \phi(x))$$
$$= h(x) \exp(\theta^\top \phi(x) - \zeta(\theta)) \quad \text{with } \zeta(\theta) = \log(Z(\theta))$$

Place a prior over θ:

$$p(\theta|a, b) \propto \eta(\theta) \exp(\theta^\top a - \zeta(\theta) b)$$

Conjugate exponential family priors

Let observations come from an exponential-family:

$$\begin{aligned} p(x|\theta) &= \frac{1}{Z(\theta)} h(x) \exp(\theta^\top \phi(x)) \\ &= h(x) \exp(\theta^\top \phi(x) - \zeta(\theta)) \quad \text{with } \zeta(\theta) = \log(Z(\theta)) \end{aligned}$$

Place a prior over θ:

$$p(\theta|a, b) \propto \eta(\theta) \exp(\theta^\top a - \zeta(\theta) b)$$

Given a set of observations $X = \{x_1, \ldots, x_N\}$

$$p(\theta|X) \propto \left(\prod_{i=1}^{N} h(x_i) \exp(\theta^\top \phi(x_i) - \zeta(\theta)) \right) \eta(\theta) \exp(\theta^\top a - \zeta(\theta) b)$$

Conjugate exponential family priors

Let observations come from an exponential-family:

$$\begin{aligned} p(x|\theta) &= \frac{1}{Z(\theta)} h(x) \exp(\theta^\top \phi(x)) \\ &= h(x) \exp(\theta^\top \phi(x) - \zeta(\theta)) \quad \text{with } \zeta(\theta) = \log(Z(\theta)) \end{aligned}$$

Place a prior over θ:

$$p(\theta|a, b) \propto \eta(\theta) \exp(\theta^\top a - \zeta(\theta) b)$$

Given a set of observations $X = \{x_1, \ldots, x_N\}$

$$\begin{aligned} p(\theta|X) &\propto \left(\prod_{i=1}^{N} h(x_i) \exp(\theta^\top \phi(x_i) - \zeta(\theta)) \right) \eta(\theta) \exp(\theta^\top a - \zeta(\theta) b) \\ &\propto \eta(\theta) \exp\left(\theta^\top \left(a + \sum_{i=1}^{N} \phi(x_i) \right) - \zeta(\theta)(b + N) \right) \end{aligned}$$

Conjugate priors (contd.)

Prior over θ: exp. fam. distribution with parameters (a, b).

Posterior: same family with parameters $(a + \sum_{i=1}^{N} \phi(x_i), b + N)$.

Rare instance where analytical expressions for posterior exists.

In most cases a simple prior quickly leads to a complicated posterior, requiring Monte Carlo methods.

Conjugate priors: Beta-Bernoulli example

Let $x \sim \text{Bern}(\pi)$, so that

$$
\begin{aligned}
p(x|\pi) &= \pi^{\mathbb{1}(x=1)}(1-\pi)^{\mathbb{1}(x=2)} \\
&= \exp\left(\mathbb{1}(x=1)\log(\pi) + (1-\mathbb{1}(x=1))\log(1-\pi)\right) \\
&= (1-\pi)\exp\left(\mathbb{1}(x=1)\log\frac{\pi}{1-\pi}\right) \\
&= \frac{1}{1+\exp(\theta)}\exp\left(\phi(x)\theta\right)
\end{aligned}
$$

Conjugate priors: Beta-Bernoulli example

Let $x \sim \text{Bern}(\pi)$, so that

$$\begin{aligned} p(x|\pi) &= \pi^{\mathbb{1}(x=1)}(1-\pi)^{\mathbb{1}(x=2)} \\ &= \exp\left(\mathbb{1}(x=1)\log(\pi) + (1-\mathbb{1}(x=1))\log(1-\pi)\right) \\ &= (1-\pi)\exp\left(\mathbb{1}(x=1)\log\frac{\pi}{1-\pi}\right) \\ &= \frac{1}{1+\exp(\theta)}\exp\left(\phi(x)\theta\right) \end{aligned}$$

This is an exponential family distrib., with
$\theta = \log\frac{\pi}{1-\pi}, \phi(x) = \mathbb{1}(x=1), h(x) = 1, Z(\theta) = (1+\exp(\theta))$.

Conjugate priors: Beta-Bernoulli example

Let $x \sim \text{Bern}(\pi)$, so that

$$\begin{aligned} p(x|\pi) &= \pi^{\mathbb{1}(x=1)}(1-\pi)^{\mathbb{1}(x=2)} \\ &= \exp\left(\mathbb{1}(x=1)\log(\pi) + (1-\mathbb{1}(x=1))\log(1-\pi)\right) \\ &= (1-\pi)\exp\left(\mathbb{1}(x=1)\log\frac{\pi}{1-\pi}\right) \\ &= \frac{1}{1+\exp(\theta)}\exp\left(\phi(x)\theta\right) \end{aligned}$$

This is an exponential family distrib., with
$\theta = \log\frac{\pi}{1-\pi}, \phi(x) = \mathbb{1}(x=1), h(x) = 1, Z(\theta) = (1+\exp(\theta))$.
Defining $\zeta(\theta) = \log Z(\theta)$ as in the previous slide,

$$p(x|\theta) = \exp\left(\phi(x)\theta - \zeta(\theta)\right)$$

Conjugate priors: Beta-Bernoulli example

If the parameter θ (or equivalently π) is unknown, Bayesian inference places a prior on it.

As before, define an exp. fam. prior with parameters $\vec{a}$:

$$p(\theta|\vec{a}) \propto \exp(a_1\theta + a_2\zeta(\theta))$$

Conjugate priors: Beta-Bernoulli example

If the parameter θ (or equivalently π) is unknown, Bayesian inference places a prior on it.

As before, define an exp. fam. prior with parameters $\vec{a}$:

$$p(\theta|\vec{a}) \propto \exp(a_1\theta + a_2\zeta(\theta))$$

Then given data $X = (x_1, \ldots, x_N)$,

$$\begin{aligned} p(\theta|\vec{a}, X) &\propto p(\theta, X|\vec{a}) \\ &\propto \exp\left(\left(a_1 + \sum_{i=1}^{N} \mathbb{1}(x_i = 1)\right)\theta + (a_2 - N)\zeta(\theta)\right) \end{aligned}$$

Conjugate priors: Beta-Bernoulli example

If the parameter θ (or equivalently π) is unknown, Bayesian inference places a prior on it.

As before, define an exp. fam. prior with parameters $\vec{a}$:

$$p(\theta|\vec{a}) \propto \exp(a_1\theta + a_2\zeta(\theta))$$

Then given data $X = (x_1, \ldots, x_N)$,

$$\begin{aligned} p(\theta|\vec{a}, X) &\propto p(\theta, X|\vec{a}) \\ &\propto \exp\left(\left(a_1 + \sum_{i=1}^{N} \mathbb{1}(x_i = 1)\right)\theta + (a_2 - N)\zeta(\theta)\right) \end{aligned}$$

Thus, the posterior is in the same family as the prior, but with updated parameters $\left(a_1 + \sum_{i=1}^{N} \mathbb{1}(x_i = 1), a_2 - N\right)$.

Conjugate priors: Beta-Bernoulli example

Looking at the prior more carefully, we see:

$$\begin{aligned} p(\theta|\vec{a}) &\propto \exp(a_1\theta + a_2\zeta(\theta)) \\ &\propto \exp\left(a_1 \log\frac{\pi}{1-\pi} + a_2\log(1-\pi)\right) \\ &\propto \pi^{a_1}(1-\pi)^{(a_2-a_1)} \\ &= \pi^{b_1-1}(1-\pi)^{(b_2-1)} \end{aligned}$$

This is just the Beta(b_1, b_2) distribution, and you can check that the posterior is Beta$\left(b_1 + \sum_{i=1}^N \mathbb{1}(x_i = 1), b_2 + \sum_{i=1}^N \mathbb{1}(x_i = 2)\right)$.

Conjugate priors: Beta-Bernoulli example

Looking at the prior more carefully, we see:

$$\begin{aligned} p(\theta|\vec{a}) &\propto \exp(a_1\theta + a_2\zeta(\theta)) \\ &\propto \exp\left(a_1 \log\frac{\pi}{1-\pi} + a_2\log(1-\pi)\right) \\ &\propto \pi^{a_1}(1-\pi)^{(a_2-a_1)} \\ &= \pi^{b_1-1}(1-\pi)^{(b_2-1)} \end{aligned}$$

This is just the Beta(b_1, b_2) distribution, and you can check that the posterior is Beta$\left(b_1 + \sum_{i=1}^N \mathbb{1}(x_i = 1), b_2 + \sum_{i=1}^N \mathbb{1}(x_i = 2)\right)$.

b_1 and b_2 are sometimes called pseudo-observations, and capture our prior beliefs: before seeing any x's our prior is as if we saw b_1 ones and b_2 twos. After seeing data, we factor actual observations into the pseudo-observations.