Beta divergence and semi-modular inference

Sitong Liu

February 2024

1 Introduction

In generalized Bayes, we update belief according to some loss function

$$\pi^l(\theta|y) = \frac{\pi(\theta)\exp\{-\eta l(\theta, y)\}}{\int \pi(\theta)\exp\{-\eta l(\theta, y)\}d\theta}, \tag{1}$$

where the parameter $\eta > 0$ calibrates the loss with the prior to accounts for the fact that $\exp(-l(\theta, y))$ is no longer constrained to integrate to 1.

If we take the loss to be the negative log-likelihood function, i.e. $l(\theta, y) = -\log f(y; \theta)$, and $\eta = 1$, then it is regular Bayes.

There are many other loss functions we can consider, such as $\beta-$divergence loss,

The $\beta-$divergence (βD) between probability densities $g(\cdot)$ and $f(\cdot)$ is defined as

$$D_B^{(\beta)}(g||f) = \frac{1}{\beta(\beta-1)}\int g^\beta dx + \frac{1}{\beta}\int f^\beta dx - \frac{1}{\beta-1}\int g f^{\beta-1}dx, \tag{2}$$

where $\beta \in \mathbb{R} \setminus \{0, 1\}$. When $\beta = 1$, $D_B^{(1)}(g(x)||f(x)) = D_{\mathrm{KL}}(g(x)||f(x))$, where D_{KL} is the Kullback-Leibler divergence defined as

$$D_{\mathrm{KL}}(g||f) = \int g \log\frac{g}{f}dx \tag{3}$$

In practice, we estimate the second term in Eq **??** with an estimate by taking $x_j \sim f(\cdot; \theta),\ \ j = 1, \ldots, m$,

$$\hat{l}_\beta(\theta) = -\frac{1}{\beta-1}\sum_{i=1}^n f(y_i; \theta)^{\beta-1} + n\frac{1}{\beta}\frac{1}{m}\sum_{j=1}^m f(x_j; \theta)^{\beta-1} \tag{4}$$

β-divergence is a special case of Bregman-divergence (BD). Let $f(x),\ \ x > 0$ be strictly convex and continuously differentiable. The functional BD is

$$\begin{aligned} D_f(p||q) &= \int f(p(x))\nu(dx) - \int f(q(x))\nu(dx) - \int f'(q(x))(p(x) - q(x))\nu(dx) \\ &= C + E_{x\sim q}\Big\{f'(q(x)) - \frac{f(q(x))}{q(x)}\Big\} - E_{y\sim p}(f'(q(y))), \end{aligned}$$

where C does not depend on q. If $f(x) = x\log(x) - x + 1$, then we get KL-divergence. If $f_\beta(x) = \frac{x^\beta - \beta x + \beta - 1}{\beta(\beta-1)}, \beta \in \mathbb{R} \setminus \{0, 1\}$, then we get β-divergence.

2 Two-Module System

We consider a two-module system, see Fig 1 for graphical representation. The first module has data Z and one parameter ϕ, while the second module has data Y and two parameters θ and ϕ. We suspect that the second module is misspecified. We consider several methods to deal with model misspecification.

We write the posterior update into two stages. We introduce an auxiliary variable $\tilde{\theta}$, which has the same distribution as θ. In the first stage, we derive the posterior of ϕ and $\tilde{\theta}$ according to

$$p(\phi, \tilde{\theta}|Z, Y) \propto \pi(\tilde{\theta}, \phi) \exp\{-\eta l_\beta(\tilde{\theta}, \phi; y)\} p(Z|\phi), \tag{5}$$

where η, β are hyper-parameters. In the second stage, we update θ using information from the first stage.

Specifically, there are a few cases that are worth discussing.

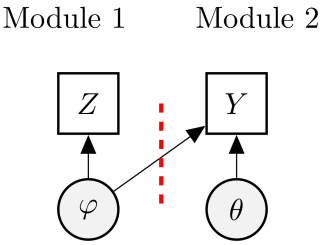


Figure 1: Graphical representation of a two-module system

2.1 Regular Bayes

If $\eta = 1$, $\beta = 1$, Eq 5 is the same as regular Bayes.

$$\begin{aligned} p(\tilde{\theta}, \phi|Y, Z) &\propto \pi(\tilde{\theta}, \phi) p(Z|\phi) p(Y|\phi, \tilde{\theta}), \\ \theta &\sim \tilde{\theta}. \end{aligned}$$

2.2 Power Posterior

If $\eta \in [0, 1]$, $\beta = 1$, Eq 5 is the same as power posterior.

$$\begin{aligned} p(\tilde{\theta}, \phi|Y, Z) &\propto \pi(\phi, \tilde{\theta}) p(Z|\phi) p(Y|\phi, \tilde{\theta})^\eta, \\ \theta &\sim \tilde{\theta}. \end{aligned}$$

2.3 Semi-Modular Inference

$\eta \in [0, 1]$, $\beta = 1$.

$$\begin{aligned} p(\tilde{\theta}, \phi|Y, Z) &\propto \pi(\phi, \tilde{\theta}) p(Z|\phi) p(Y|\phi, \tilde{\theta})^\eta, \\ \theta &\sim \pi(\theta|Y, \phi), \end{aligned}$$

where $\pi(\theta|Y, \phi) \propto p(Y|\phi, \theta)\pi(\theta|\phi)$.

2.4 β-loss Posterior

$\eta = 1,\ \ \beta \in \mathbb{R} \setminus \{0, 1\}$.

$$p(\phi, \tilde{\theta}|Z, Y) \propto \pi(\phi, \tilde{\theta}) \exp\{-l_\beta(\tilde{\theta}, \phi; y)\} p(Z|\phi),$$
$$\theta \sim \tilde{\theta}.$$

2.5 SMI-β Posterior

$\eta > 0,\ \ \beta \in \mathbb{R} \setminus \{0, 1\}$

$$p(\tilde{\theta}, \phi|Y, Z) \propto \pi(\phi, \tilde{\theta}) p(Z|\phi) \exp\{-\eta l_\beta(\tilde{\theta}, \phi; y)\},$$
$$\theta \sim \tilde{\theta}.$$

2.6 β-β' Posterior

$\beta, \beta' \in \mathbb{R} \setminus \{0, 1\}$.

$$p(\phi, \tilde{\theta}|Z, Y) \propto \pi(\phi, \tilde{\theta}) \exp\{-l_\beta(\tilde{\theta}, \phi; y)\} p(Z|\phi),$$
$$\theta \sim \pi(\theta|\phi) \exp\{-l_{\beta'}(\phi, \theta; y)\}.$$

2.7 Power-β Posterior

$\eta \in [0, 1],\ \ \beta \in \mathbb{R} \setminus \{0, 1\}$

$$p(\tilde{\theta}, \phi|Y, Z) \propto \pi(\phi, \tilde{\theta}) p(Z|\phi) p(Y|\phi, \tilde{\theta})^\eta,$$
$$\theta \sim \pi(\theta|\phi) \exp\{-l_\beta(\phi, \theta; y)\}$$

2.8 Two stage β-loss Posterior

$\eta = 1,\ \ , \beta \in \mathbb{R} \setminus \{0, 1\}$

$$p(\phi, \tilde{\theta}|Z, Y) \propto \pi(\phi, \tilde{\theta}) \exp\{-l_\beta(\tilde{\theta}, \phi; y)\} p(Z|\phi),$$
$$\theta \sim \pi(\theta|Y, \phi)$$

2.9 Two stage SMI-β Posterior

$\eta > 0,\ \ \beta \in \mathbb{R} \setminus \{0, 1\}$

$$p(\tilde{\theta}, \phi|Y, Z) \propto \pi(\phi, \tilde{\theta}) p(Z|\phi) \exp\{-\eta l_\beta(\tilde{\theta}, \phi; y)\},$$
$$\theta \sim \pi(\theta|Y, \phi).$$

2.10 Two stage SMI-$\beta - \beta'$ Posterior

$\eta > 0,\ \ \beta, \beta' \in \mathbb{R} \setminus \{0, 1\}$.

$$p(\tilde{\theta}, \phi|Y, Z) \propto \pi(\phi, \tilde{\theta}) p(Z|\phi) \exp\{-\eta l_\beta(\tilde{\theta}, \phi; y)\},$$
$$\theta \sim \pi(\theta|\phi) \exp\{-l_{\beta'}(\phi, \theta; y)\}.$$

The above ten cases are concluded in Table 1.

First stage	Second stage	Type
$\tilde{\theta}, \phi \sim \pi(\cdot\|y,z)$	$\theta = \tilde{\theta}$	Regular Bayes
$\tilde{\theta}, \phi \sim \pi_{pow,\eta}(\cdot\|y,z)$	$\theta = \tilde{\theta}$	Power Posterior
$\tilde{\theta}, \phi \sim \pi_{\beta}(\cdot\|y,z)$	$\theta = \tilde{\theta}$	β-loss Posterior
$\tilde{\theta}, \phi \sim \pi_{pow,\eta}(\cdot\|y,z)$	$\theta \sim \pi(\cdot\|y,\phi)$	SMI
$\tilde{\theta}, \phi \sim \pi_{\beta,\eta}(\cdot\|y,z)$	$\theta = \tilde{\theta}$	SMI-β
$\tilde{\theta}, \phi \sim \pi_{\beta}(\cdot\|y,z)$	$\theta \sim \pi_{\beta'}(\cdot\|y,\phi)$	$\beta - \beta'$ Posterior
$\tilde{\theta}, \phi \sim \pi_{pow,\eta}(\cdot\|y,z)$	$\theta \sim \pi_{\beta}(\cdot\|y,\phi)$	Power-β
$\tilde{\theta}, \phi \sim \pi_{\beta}(\cdot\|y,z)$	$\theta \sim \pi(\cdot\|y,\phi)$	Two stage β-loss
$\tilde{\theta}, \phi \sim \pi_{\beta,\eta}(\cdot\|y,z)$	$\theta \sim \pi(\cdot\|y,\phi)$	Two stage SMI-β Posterior
$\tilde{\theta}, \phi \sim \pi_{\beta,\eta}(\cdot\|y,z)$	$\theta \sim \pi_{\beta'}(\cdot\|y,\phi)$	Two stage SMI-$\beta - \beta'$ Posterior

Table 1: Summary of distributions in stages 1 & 2 for various versions of modular inference

3 Example

3.1 Biased data

We consider a synthetic example where the misspecification comes from a poorly chosen prior. Suppose we have two datasets informing an unknown parameter ϕ. The first is a "reliable" small sample $Z = (Z_1, \ldots, Z_n)$, $z_i \sim N(\phi, \sigma_z^2)$, iid for $i = 1, \ldots, n$, with σ_z known. The second is a larger sample $Y = (Y_1, \ldots, Y_m)$, $y_j \sim N(\phi + \theta, \sigma_y^2)$, iid for $j = 1, \ldots, m$, with σ_y known. The "bias" θ is unknown.

We take $n = 25$ and $m = 50$. The true parameter values are $\phi^\star = 0, \theta^\star = 1$, and we know $\sigma_z = 2, \sigma_y = 1$. We assign a constant prior for ϕ and $\pi(\theta) = N(0, \sigma_\theta^2)$, where we choose $\sigma_\theta = 0.33$. We want to check whether inference using β-loss is robust to outliers. We intentionally construct dataset with outliers, i.e., the true data generating process for y is $y_j \sim N(\phi^\star + \theta^\star, \sigma_y')$, iid for $j = 1, \ldots, m$, where $\sigma_y' > \sigma_y$.

We use different belief updates to update the posterior and compare the results in terms of expected log pointwise predictive density (elpd).

As mentioned in Carmona and Nicholls (2020),

$$elpd(\eta) = \int\int p^\star(z,y) \log p_{smi,\eta}(z,y|Z,Y) dz dy,$$

where $p^\star$ is the true data-generating process and

$$p_{smi,\eta}(z,y|Z,Y) = \int\int p(z,y|\phi,\theta) p_{smi,\eta}(\phi,\theta|Y,Z) d\phi d\theta.$$

Nicholls et al. (2022) gives the detail to calculate elpd in terms of δ-SMI. For β-loss, we can calculate it explicitly.

$$\begin{aligned}
l_\beta(y, f(\cdot;\theta,\phi)) &= -\frac{1}{\beta-1} N(y;\phi+\theta,\sigma_y^2)^{\beta-1} + \frac{1}{\beta}\int N(z;\phi+\theta,\sigma_y^2)^\beta dz \\
&= -\frac{1}{\beta-1} N(y;\phi+\theta,\sigma_y^2)^{\beta-1} + \frac{1}{\beta}\int \Big\{\frac{1}{\sqrt{2\pi}\sigma_y}\exp\Big(-\frac{(z-(\phi+\theta))^2}{2\sigma_y^2}\Big)\Big\}^\beta dz \\
&= -\frac{1}{\beta-1} N(y;\phi+\theta,\sigma_y^2)^{\beta-1} + \frac{1}{\beta}(2\pi)^{-\frac{\beta}{2}}\sigma_y^{-\beta}\cdot(\sqrt{2\pi}\frac{\sigma_y}{\sqrt{\beta}})\int \frac{1}{\sqrt{2\pi}\frac{\sigma_y}{\sqrt{\beta}}}\exp\Big(-\frac{(z-(\phi+\theta))^2}{2(\frac{\sigma_y}{\sqrt{\beta}})^2}\Big) dz \\
&= -\frac{1}{\beta-1} N(y;\phi+\theta,\sigma_y^2)^{\beta-1} + \beta^{\frac{1}{2}}(2\pi)^{-\frac{\beta-1}{2}}(\sigma_y)^{-\beta+1}.
\end{aligned}$$

For elpd in β-loss posterior, we have

$$\begin{aligned}
elpd(\beta) &= \int\int p^\star(z,y) \log p_\beta(z,y|Z,Y) dz dy, \\
p_\beta(z,y|Z,Y) &= \int\int p(z,y|\phi,\theta) p_\beta(\phi,\theta|Y,Z) d\phi d\theta.
\end{aligned}$$

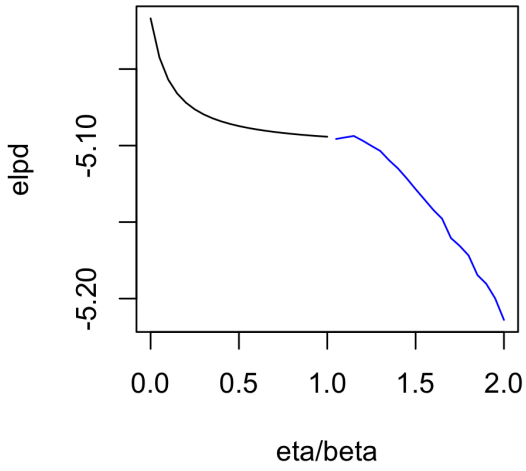


Figure 2: Elpd of both y and z modules for biased data example with $\sigma_y' = \sigma_y + 1$. The black line is the exact elpd of SMI with $\eta \in [0, 1]$ and the blue line is beta posterior with $\beta \in [1.05, 2.0]$.

Case	Optimal	ELPD
SMI (2.3)	$\eta^\star = 0$	-5.281
β-loss posterior (2.4)	$\beta^\star = 1.15$	-5.323
SMI-β (2.5)	$\eta^\star = 0.85$, $\beta^\star$=1.15	-5.316
$\beta - \beta'$ (2.6)	$\beta^\star = 2.4$, $\beta'^\star = 1.2$	-5.335
Power-β (2.7)	$\eta^\star = 0$, $\beta^\star = 1.05$	-5.280
Two stage β-loss (2.8)	$\beta^\star = 1.85$	-5.339
Two stage SMI-β (2.9)	$\beta^\star = 1.1$, $\eta^\star = 0$	-5.272

Table 2: The optimal parameter value of various versions of modular inference

As we cannot express the β-loss posterior for ϕ, θ in explicit forms, we can use Monte Carlo samples to estimate elpd instead. We simulate Monte Carlo samples $\{(z_0^{(i)}, y_0^{(i)})\}_{i=1}^N$ from the true generative distribution $p^\star$, and for each sample pair $(z_0^{(i)}, y_0^{(i)})$, we estimate $p_\beta(z_0^{(i)}, y_0^{(i)}|Z, Y)$ by taking Monte Carlo samples of ϕ, θ from the β-loss posterior $p_\beta(\phi, \theta|Y, Z)$ and evaluate these values in $p_\beta(z_0^{(i)}, y_0^{(i)}|Z, Y)$.

From Fig 2, SMI suggests taking $\eta = 0$, which is the cut, while β-loss posterior with $\beta = 1.15$ gives the maximal elpd, which is close to Bayes.

We repeat the procedures above 100 times, and each time we generate dataset (Z, Y) independently. We compare the difference of maximal elpd between β-loss posterior and SMI in Fig 4. As we can see, the differences are always negative, so SMI performs better in prediction than β-loss posterior.

We can also combine η, β as discussed in section 2.5. We take $\beta \in [1.05, 2]$ and for each β, we let its associated $\eta \in [0, \beta]$. We plot the maximal elpd of each β in Fig **??**. In this example, $(\beta^\star, \eta^\star) = (1.15, 0.85)$ gives the maximal elpd=-5.316.

We summarize the optimal cases of methods discussed in section 2 in table3.1

We also calculate posterior mean squared error (PMSE) for SMI and β-loss posterior. For SMI, we can exactly calculate PMSE.

We can also use Gaussian quadrature to estimate ELPD as the errors are negligible in Monte Carlo simulation.

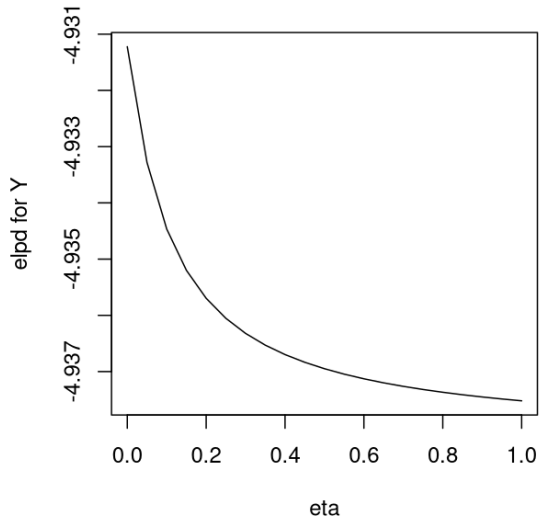


Figure 3: Elpd for Y module for biased data example, which prefers the cut.

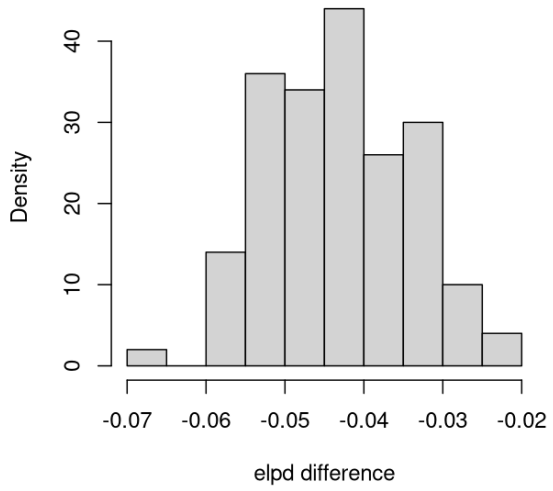


Figure 4: Histogram of the difference in maximal elpd between β-loss and SMI.

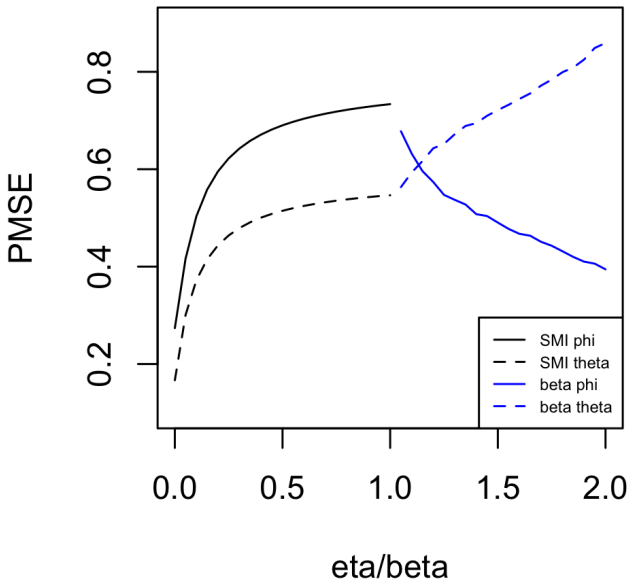


Figure 5: (Normal biased data) PMSE for ϕ (solid) and θ (dashed) of theoretical SMI (black) and β-loss posterior (blue)

3.1.1 Safe Bayes

Apart from elpd, there are many other methods to choose η and β. We now consider Safe Bayes (Grunwald and Van Ommen, 2017). We keep Z-module fixed, and add one Y data each time.

For R-log loss,

$$\begin{aligned} r &:= l_{\Pi|Z,Y^{i-1},\eta}(Y_i) \\ &= E_{\theta\sim\Pi|Z,Y^{i-1},\eta}[l_\theta(Y_i)] \\ &= E_{(\theta,\phi)\sim\Pi|Z,Y^{i-1},\eta}[-\log p(Y_i|\theta,\phi)] \\ &= \frac{1}{2}\log(2\pi) + \log\sigma_{y_\eta} + \frac{(y_i-(\theta_\eta+\phi_\eta))^2}{2\sigma_{y_\eta}^2} + \frac{Var(\theta+\phi|Z,Y^{i-1})}{2\sigma_{y_\eta}^2} \\ &= -\log N(y_i;\theta_\eta+\phi_\eta,\sigma_{y_\eta}^2) + \frac{Var(\theta+\phi|Z,Y^{i-1},\eta)}{2\sigma_{y_\eta}^2}, \end{aligned}$$

where θ_y, ϕ_y are posterior mean for θ and ϕ respectively, and $\sigma_{y_\eta}^2 = Var(\theta+\phi|Z,Y^{i-1},\eta)+\sigma_y^2$.

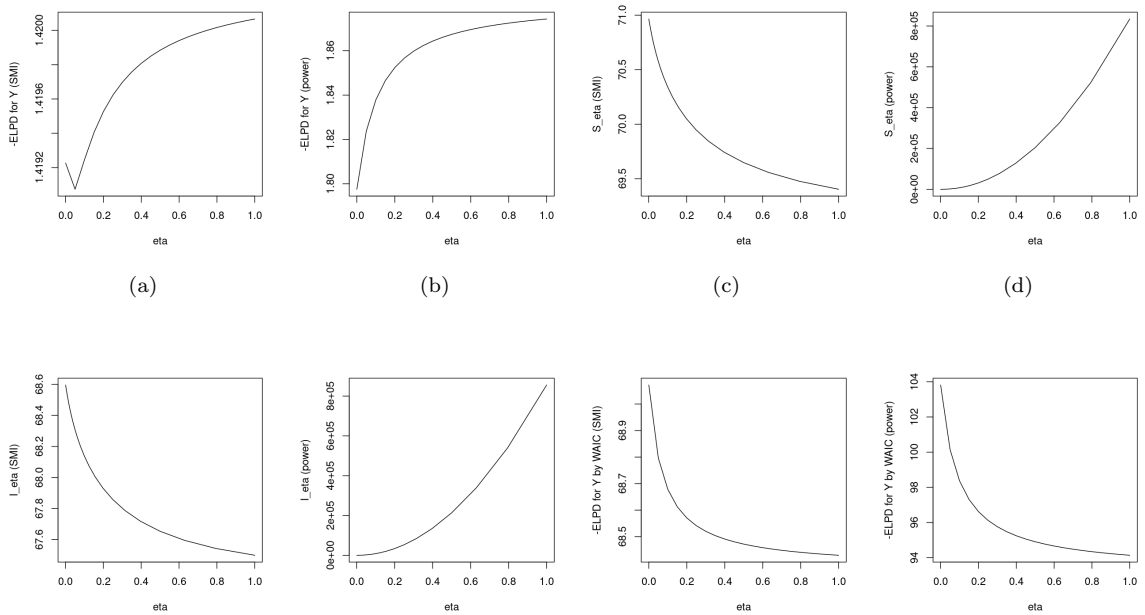
For I-log loss, where we use posterior mean,

$$\begin{aligned} i &:= E_{(\theta,\phi)\sim\Pi|Z,Y^{i-1},\eta}[-\log p(Y_i|\theta_\eta,\phi_\eta)] \\ &= -\log N(y_i;\theta_\eta+\phi_\eta,\sigma_y^2) \end{aligned}$$

I choose $\kappa_{\text{STEP}} = \frac{1}{3}$ and $\kappa_{\max} = 10$. Fig 12 shows how s_n changes with η and we choose $\hat{\eta} = 0$, the cut model. We also consider use only the first stage of η-SMI, which is equivalent to power posterior, which suggest choose $\eta \approx 0.25$.

3.1.2 Compare Safe Bayes and ELPD

We now compare Safe Bayes and ELPD on different datasets, using SMI and power posterior.



(a) (b) (c) (d)

(e) (f) (g)

Figure 6: Biased normal data with $\sigma_y = 1$. (a) -ELPD for SMI (b) -ELPD for power posterior (c) R-Safe Bayes for SMI (d) R-Safe Bayes for power posterior (e) I-Safe Bayes for SMI (f) I-Safe Bayes for power posterior (g) -ELPD estimated by WAIC for SMI (h) -ELPD estimated by WAIC for power posterior

SMI/pow	Safe Bayes (R)	Safe Bayes (I)	ELPD	WAIC
Biased normal, $\sigma_y = 1$	Bayes/Cut	Bayes/Cut	$\eta = 0.05$/Cut	Bayes/Bayes
Biased normal, $\sigma_y = 2$	$\eta = 0.8$/Cut	Bayes/Cut	Cut/Cut	Bayes/Bayes
Mixture	Cut/Cut	Cut/Cut	Bayes/Cut	Bayes/Bayes

Table 3: Summary of the optimal η for SMI and power posterior selected by Safe Bayes and ELPD.

We set seed equal to 1 when generating Z, Y for all datasets we use.

For biased data with $\sigma_y = 1$, ELPD choose $\eta = 0.05$ for SMI. However, if we change seed, ELPD choose $\eta = 1$. ELPD choose $\eta - 0.05$ for power posterior. For Safe Bayes, it chooses $\eta = 1$ for SMI (still choose $\eta = 1$ if we change seed), and chooses $\eta \approx 0.2$ for power posterior.

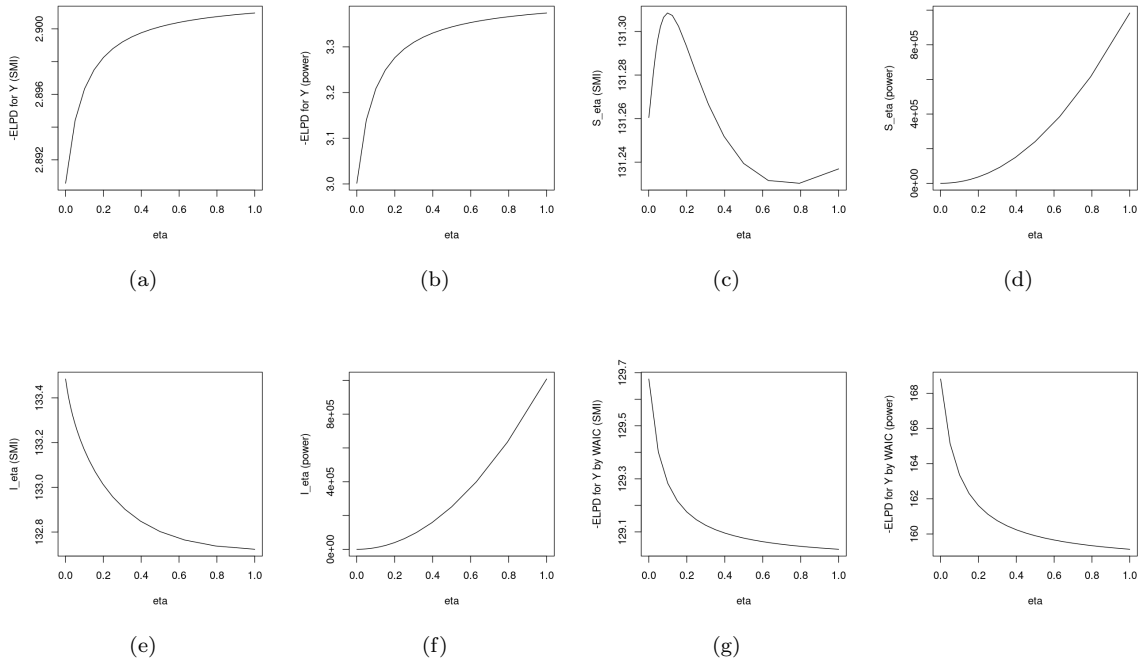


Figure 7: Biased normal data with $\sigma_y = 2$. (a) -ELPD for SMI (b) -ELPD for power posterior (c) R-Safe Bayes for SMI (d) R-Safe Bayes for power posterior (e) I-Safe Bayes for SMI (f) I-Safe Bayes for power posterior (g) -ELPD estimated by WAIC for SMI (h) -ELPD estimated by WAIC for power posterior

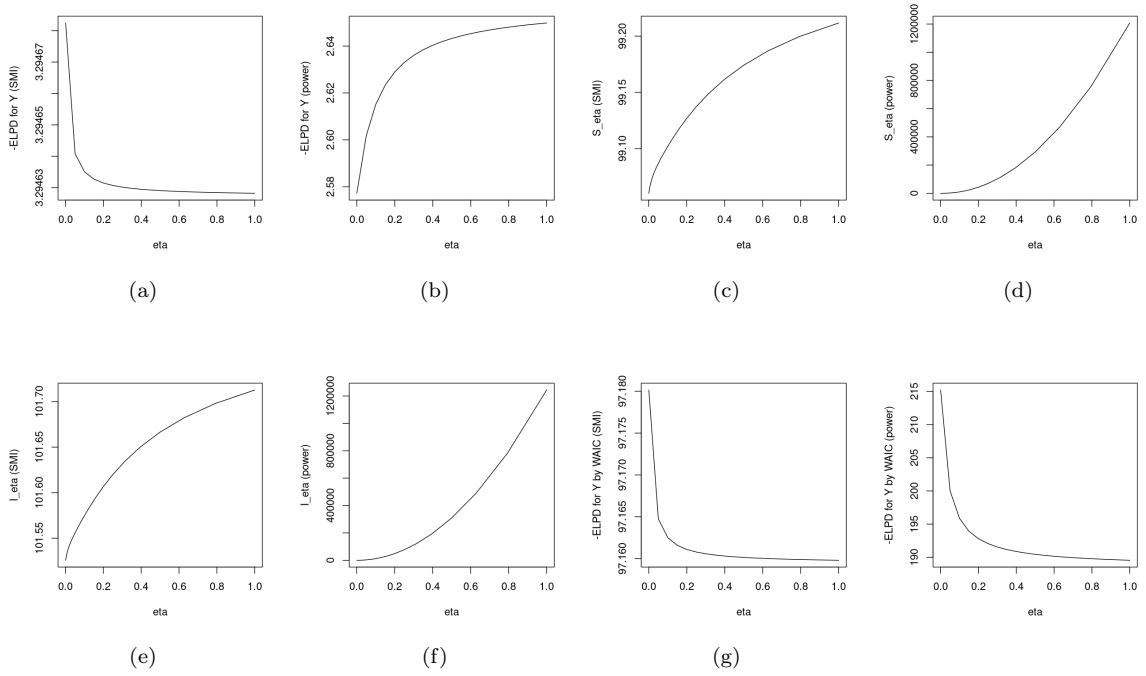
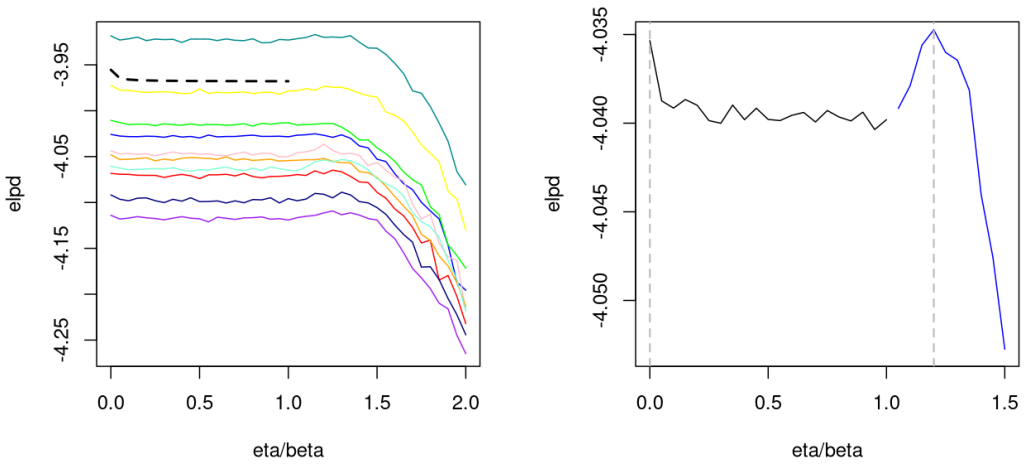


Figure 8: Mixture Gaussian data. (a) -ELPD for SMI (b) -ELPD for power posterior (c) R-Safe Bayes for SMI (d) R-Safe Bayes for power posterior (e) I-Safe Bayes for SMI (f) I-Safe Bayes for power posterior (g) -ELPD estimated by WAIC for SMI (h) -ELPD estimated by WAIC for power posterior



(a) (b)

Figure 9: Elpd for Gaussian mixture model using β-loss and SMI. (a) Each colored line is the Monte Carlo estimate of elpd, where x-axis value smaller than 1 is SMI and greater than 1 is β-loss posterior. The black dashed line is the theoretical elpd value for SMI. (b) Average of ten repeats in (a). The black line is for SMI and blue for β-loss posterior. The vertical grey dotted lines are at x=0 and x=1.2, corresponding to η, β with maximal elpd value

3.2 Biased data with mixture of Gaussian model

We consider a similar dataset as in section 3.1, and the difference is that the true generating distribution for Y is a Gaussian mixture, i.e., $y_j \sim 0.9 \cdot N(\phi^\star + \theta^\star, 1) + 0.1 \cdot N(4, 1)$, iid for $j = 1, \ldots, m$. The prior for ϕ and θ are both standard normal distributions. The likelihood for the data is the same as in section 3.1.

We use β-loss and SMI for inference. We consider $\beta \in (1, 2)$ and $\eta \in [0, 1]$. We use elpd as the criterion to select the optimal β and η. We repeat the Monte Carlo estimation ten times, and there are seven out of ten times that the maximal elpd of β-loss posterior ($\beta^\star \approx 1.2$) is greater than the maximal elpd of SMI ($\eta^\star = 0$), see Figure 9. The average best performance of β-loss posterior beats SMI.

There are several other methods to choose the optimal β and η.

3.2.1 Posterior Mean Squared Error

As we know the true parameter value in this example, we can calculate posterior mean squared error (PMSE), i.e. $E[(\phi - \phi^\star)^2 | Y, Z]$ and $E[(\theta - \theta^\star)^2 | Y, Z]$. Figure **??** are PMSE for SMI and β-loss posterior.

We can derive a theoretical formula to calculate PMSE for SMI.

$$
\begin{aligned}
E[(\phi - \phi^\star)^2 | Y, Z] &= E(\phi^2 | Y, Z), \quad (\phi^\star = 0) \\
&= Var(\phi | Y, Z) + E(\phi | Y, Z)^2 \\
&= \frac{\lambda \sigma_z^2}{n} + [\lambda \bar{Z} + (1 - \lambda)\bar{Y}]^2, \quad \lambda = \frac{n/\sigma_z^2}{n/\sigma_z^2 + m/(\sigma_y^2/\eta + m\sigma_\theta^2) + 1/\sigma_\phi^2}, \\
E[(\theta - \theta^\star)^2 | Y, Z] &= E[(\theta - 1)^2 | Y, Z], \quad \theta^\star = 1, \\
&= Var(\theta | Y, Z) + E(\theta | Y, Z)^2 - 2E(\theta | Y, Z) + 1
\end{aligned}
$$

Question: I only know $p(\theta | Y, \phi)$ and posterior mean of θ depends on ϕ. How can I integrate ϕ out? Answer: See "PMSE SMI derivation.tex" for derivation, but I guess something goes wrong as the result is strange according to the formula I derive.

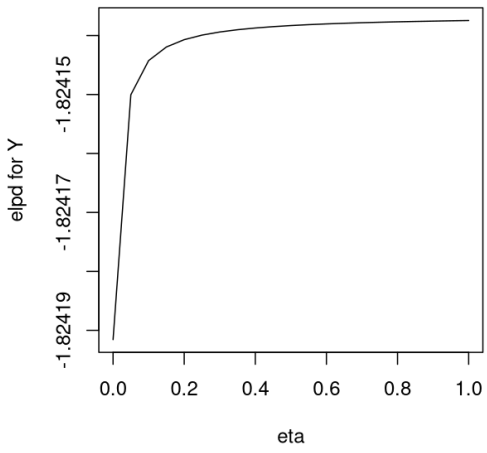


Figure 10: ELPD for Y module, mixture example.

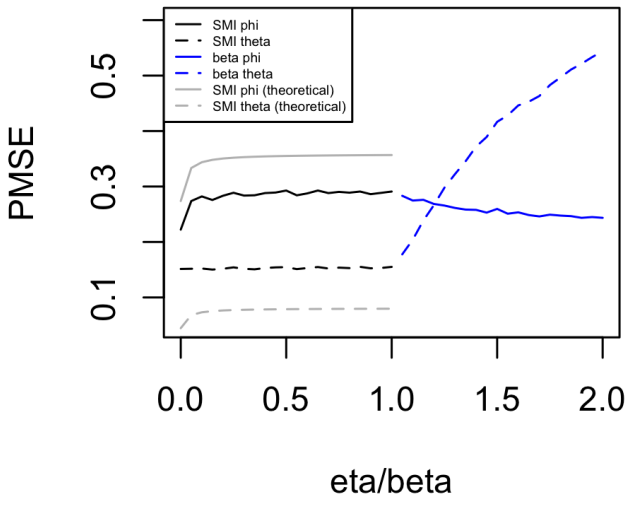
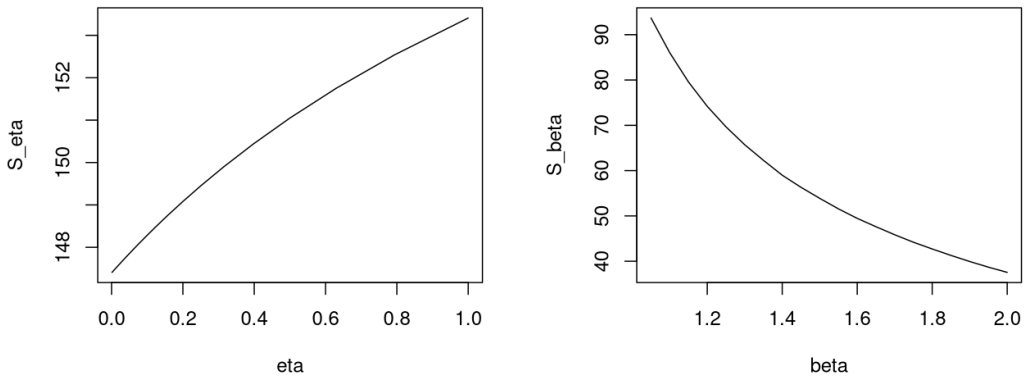


Figure 11: (Gaussian mixture example) PMSE for ϕ (solid) and θ (dashed) of SMI (black is Monte Carlo estimates and grey is theoretical value) and β-loss posterior (blue)

Case	Optimal	ELPD
SMI (2.3)	$\eta^\star = 0$	-4.068
β-loss posterior (2.4)	$\beta^\star = 1.25$	-4.065
Power-β (2.7)	$\eta^\star = 0,\ \ \beta^\star = 1.05$	-4.084
Two stage SMI-β (2.9)	$\beta^\star = 1.05,\ \ \eta^\star = 0$	

Table 4: The optimal parameter value of various versions of modular inference for mixture Gaussian example.



(a)

Figure 12: Safe Bayes to choose η in η-SMI and β in β-loss posterior. For η-SMI, we choose $\hat{\eta} = 0$, the cut model. For β-loss, we choose $\hat{\beta}$=2.

Apart from SMI and β-loss posterior, we also consider other combinations of η and β. Looking at Table 3.1, which contains all kinds of combinations of η and β to analyze biased data example, SMI, Power-β, and two stage SMI-β performs best, and they all suggest the cut. We now test these three methods on mixture Gaussian data.

3.2.2 Safe Bayes

3.3 Epidemiological data

We use an epidemiological dataset to compare these methods. The model has two modules: in each population $i = 1, \ldots, 13$, a Poisson response for the number of cancer cases Y_i in T_i women-years of followup, and a Binomial model for the number Z_i of women infected with HPV in a sample of size N_i from the ith population,

$$\begin{aligned} Y_i &\sim Poisson(\mu_i), \\ \mu_i &= T_i \exp(\theta_1 + \theta_2 \phi_i), \\ Z_i &\sim Binomial(N_i, \phi_i). \end{aligned}$$

For the priors,

$$\begin{aligned} \phi_i &\sim Beta(c_1^i, c_2^i), \;\; i = 1, \ldots, 13, \\ \theta_1 &\sim N(m, s^2), \\ \theta_2 &\sim Gamma(g_1, g_2). \end{aligned}$$

We choose the hyperparameters c_1, c_2, m, s, g_1, g_2 at values selected by Carmona and Nicholls (2020), with $c_1^i = 0.1, c_2^i = 3, i = 1 \ldots, 13$ and $m = 0, s = 100, g_1 = 1, g_2 = 0.1$.

3.3.1 Beta Loss

The beta loss takes the form

$$l_\beta(y_i, f(\cdot; \theta_1, \theta_2, \phi_i)) = -\frac{1}{\beta - 1} f(y_i; \theta_1, \theta_2, \phi_i)^{\beta - 1} + \frac{1}{\beta} \int f(z; \theta_1, \theta_2, \phi_i)^\beta dz \tag{6}$$

To estimate $\int f(z; \theta_1, \theta_2, \phi)^\beta dz$, we can directly use the sum $\sum_{z=0,1,2,\ldots} f(z; \theta_1, \theta_2, \phi)^\beta$, as in our example, the Poisson distribution is discrete. We can ignore terms smaller than 10^{-16}.

The beta posterior is

$$\pi_\beta(\theta_1, \theta_2, \phi) \propto \prod_{i=1}^{13} p(z_i | \phi_i) \exp\{-\sum_{i=1}^{13} l_\beta(y_i, f(\cdot; \theta_1, \theta_2, \phi_i)\} \pi(\theta_1) \pi(\theta_2) \prod_{i=1}^{13} \pi(\phi_i).$$

We take $\theta_1 = -2, \theta_2 = 20$ and take ϕ to be its prior mean, and take $\mu = \bar{T} \exp(\theta_1 + \theta_2 \phi)$. We plot β-loss in 6 for y varying from 0 to 1000 with different β values in Fig 13. β-loss are almost zero except around μ. As β increases, β-loss tends to be flatter. With the same θ_1, θ_2, ϕ values, and with fixed y values as in the original data, we plot β-loss with different β varying from 1.05 to 2 in Fig 14

3.4 Linear model selection

This example comes from Grünwald and Van Ommen (2017).

$$\begin{aligned} Y_i &= \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \\ \epsilon_i &\sim N(0, \sigma^2), \end{aligned}$$

X_{ij} are polynomial function of S_i and $S_i \in [-1, 1]$ uniformly i.i.d,

$$\begin{aligned} s_i &\sim U(-1, 1) \\ x_{ij} &= s_i^j, \;\; j = 1, \ldots, p \end{aligned}$$

For the true data generating process, (X_i, Y_i) is set to be $(0, 0)$ with $\frac{1}{2}$ probability. This is where model misspecification comes from.

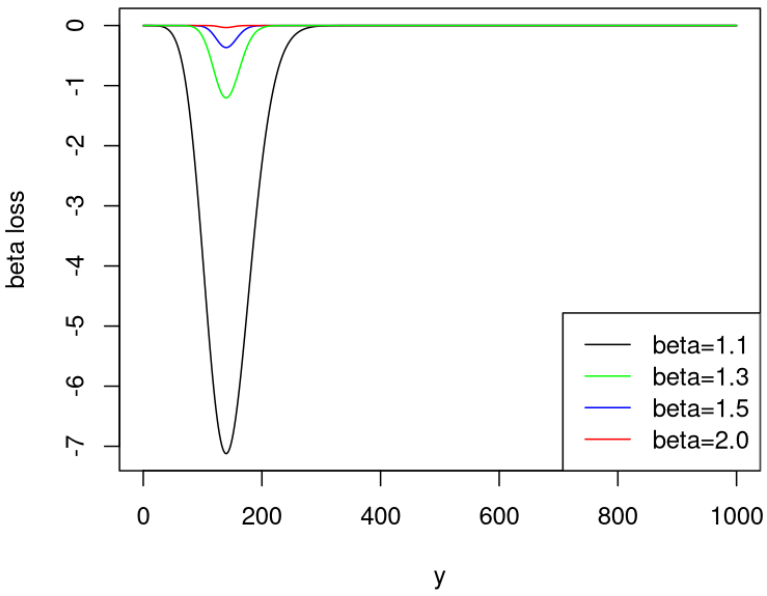


Figure 13: β-loss as a function of y with different β values.

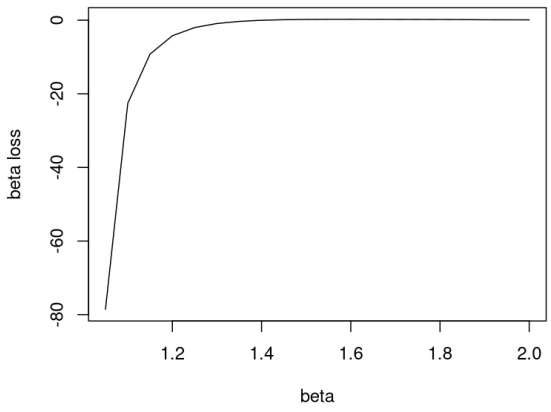


Figure 14: β-loss with different β values and fixed y.

Let $\mathbf{X}_n = (x_1^T, \ldots, x_n^T)^T$ be the design matrix. For the priors,

$$\begin{aligned}\beta|\sigma^2 &\sim N(\bar{\beta}_0, \sigma^2\Sigma_0)\\ \sigma^2 &\sim Inv-Gamma(\sigma^2|a_0, b_0) = \sigma^{-2(a+1)}e^{-b/\sigma^2}b^a/\Gamma(a)\end{aligned}$$

then the generalized posterior on β is

$$\begin{aligned}\beta_{n,\eta} &\sim N(\bar{\beta}_{n,\eta}, \sigma^2\Sigma_{n,\eta}),\\ \bar{\beta}_{n,\eta} &= \Sigma_{n,\eta}(\Sigma_0^{-1}\bar{\beta}_0 + \eta\mathbf{X}_n^T y^n),\\ \Sigma_{n,\eta} &= (\Sigma_0^{-1} + \eta\mathbf{X}_n^T\mathbf{X}_n)^{-1}\end{aligned}$$

The general posterior for σ^2 is

$$\begin{aligned}\pi(\sigma^2|z^n, p, \eta) &= Inv-Gamma(\sigma^2|a_{n,\eta}, b_{n,\eta}),\\ a_{n,\eta} &= a_0 + \eta n/2\\ b_{n,\eta} &= b_0 + \frac{\eta}{2}\sum_{i=1}^{n}(y_i - x_i\bar{\beta}_{n,\eta})^2\end{aligned}$$

The posterior expectation of σ^2 can be calculated as

$$\bar{\sigma}^2_{n,\eta} := \frac{b_{n,\eta}}{a_{n,\eta} - 1}.$$

In experiment, they set $\bar{\beta}_0 = \mathbf{0}$, $\Sigma_0 = \mathbf{I}_{p+1}$, $a_0 = 1$, $b_0 = 40$.
They use a fat-tailed prior on the models

$$\pi(p) \propto \frac{1}{(p+2)(\log(p+2))^2}$$

As given in the supplement, the I-log-loss is

$$\mathbf{E}_{\beta\sim\Pi|z^i,p,\eta}[-\log f(y_{i+1}|x_{i+1}, \beta, \sigma^2)] = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y_{i+1} - x_{i+1}\bar{\beta}_{i,\eta})^2$$

The R-log-loss is

$$\mathbf{E}_{\sigma^2,\beta\sim\Pi|z^i,p,\eta}[-\log f(y_{i+1}|x_{i+1}, \beta, \sigma^2)] = \frac{1}{2}log2\pi\bar{\sigma}^2_{i,\eta} + \frac{1}{2}\frac{(y_{i+1} - x_{i+1}\bar{\beta}_{i,\eta})^2}{\bar{\sigma}^2_{i,\eta}} + \frac{1}{2}x_{i+1}\Sigma_{i,\eta}x_{i+1}^T + r(i,\eta)$$

$r(i,\eta)$ is a remainder function which is $O(1/i)$ whenever $\sum_{i=1}^{n}(y_i - x_i\beta_{n,\eta})^2$ increases linearly in n.

References