Chapter 10
Bayesian Philosophy

10.1 Introduction

Up to now… **Classical Approach**: assumes θ is *deterministic*

This has a few ramifications:

- Variance of the estimate could depend on θ
- In Monte Carlo simulations:
  - *M* runs done at the *same* θ,
  - must do *M* runs at each θ of interest
  - averaging done over data
  - *no averaging over* θ *values*

$E\{\}$ is w.r.t. $p(\mathbf{x};\theta)$

Bayesian Approach: assumes θ is *random* with pdf $p(\theta)$

This has a few ramifications:

- Variance of the estimate CAN'T depend on θ
- In Monte Carlo simulations:
  - *each* run done at a *randomly* chosen θ,
  - averaging done over data *AND* over θ values

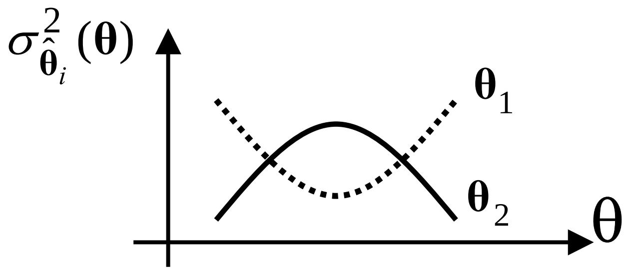
$E\{\}$ is w.r.t. $p(\mathbf{x},\theta)$

joint pdf

Why Choose Bayesian?

1. Sometimes we have prior knowledge on θ ⇒ some values are more likely than others

2. Useful when the classical MVU estimator does not exist because of nonuniformity of minimal variance



3. To combat the "signal estimation problem"… estimate signal **s**

$$\mathbf{x} = \mathbf{s} + \mathbf{w}$$

If **s** is deterministic and is the parameter to estimate, then **H** = **I**

Classical Solution: $\hat{\mathbf{s}} = \left(\mathbf{I}^T\mathbf{I}\right)^{-1}\mathbf{I}^T\mathbf{x} = \mathbf{x}$

Signal Estimate is the data itself!!!

The Wiener filter is a Bayesian method to combat this!!

10.3 Prior Knowledge and Estimation

Bayesian Data Model:

- Parameter is "chosen" randomly w/ known "prior PDF"
- Then data set is collected
- Estimate value chosen for parameter

This is what you know ahead of time about the parameter.

Every time you collect data, the parameter has a different value, but some values may be more likely to occur than others

This is how you think about it mathematically and how you run simulations to test it.

Ex. of Bayesian Viewpoint: Emitter Location

Emitters are where they are and don't randomly jump around each time you collect data. **So why the Bayesian model?**

(At least) Three Reasons

1. You may know from maps, intelligence data, other sensors, etc. that certain locations are more likely to have emitters
   - Emitters likely at airfields, unlikely in the middle of a lake
2. Recall Classical Method: Parm Est. Variance often depends on parameter
   - It is often desirable (e.g. marketing) to have a single number that measures accuracy.
3. Classical Methods try to give an estimator that gives low variance at *each* θ value. However, this could give large variance where emitters are likely and low variance where they are unlikely.

Bayesian Criteria Depend on Joint PDF

There are several different optimization criteria within the Bayesian framework. The most widely used is…

Minimize the Bayesian MSE: $Bmse(\hat{\theta}) = E\left\{(\theta - \hat{\theta})^2\right\}$

Take E{} w.r.t. joint pdf of **x** and θ

$$= \iint [\theta - \hat{\theta}(\mathbf{x})]^2 \, p(\mathbf{x},\theta) \, d\mathbf{x} \, d\theta$$

Can Not Depend on θ

Joint pdf of **x** and θ

To see the difference… compare to the Classical MSE:

$$mse(\hat{\theta}) = E\left\{(\theta - \hat{\theta})^2\right\}$$

$$= \int [\theta - \hat{\theta}(\mathbf{x})]^2 \, p(\mathbf{x};\theta) \, d\mathbf{x}$$

Can Depend on θ

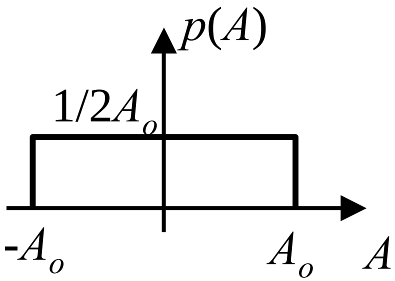
pdf of **x** parameterized by θ

Ex. Bayesian for DC Level

Same as before… $x[n] = A + w[n]$

Zero-Mean White Gaussian

But here we use the following model:
- that A is random w/ uniform pdf
- RVs A and $w[n]$ are independent of each other



Now we want to find the estimator function that maps data $\mathbf{x}$ into the estimate of A that minimizes Bayesian MSE:

$$Bmse(\hat{A}) = \iint [A - \hat{A}]^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA$$

$$= \int \underbrace{\left[\int [A - \hat{A}]^2 p(A \mid \mathbf{x})\, dA\right]}_{} p(\mathbf{x})\, d\mathbf{x}$$

Now use… $p(\mathbf{x},A) = p(A|\mathbf{x})p(\mathbf{x})$

Minimize this for <u>each</u> $\mathbf{x}$ value

This works because $p(\mathbf{x}) \geq 0$

So… fix $\mathbf{x}$, take its partial derivative, set to 0

Finding the Partial Derivative gives:

$$\frac{\partial}{\partial \hat{A}} \int [A-\hat{A}]^2 p(A|\mathbf{x})dA = \int \frac{\partial [A-\hat{A}]^2}{\partial \hat{A}} p(A|\mathbf{x})dA$$

$$= \int -2[A-\hat{A}]p(A|\mathbf{x})dA$$

$$= -2\int Ap(A|\mathbf{x})dA + 2\hat{A}\underbrace{\int p(A|\mathbf{x})dA}_{=1}$$

Setting this equal to zero and solving gives:

$$\boxed{\begin{aligned}\hat{A} &= \int Ap(A|\mathbf{x})dA \\ &= E\{A|\mathbf{x}\}\end{aligned}}$$

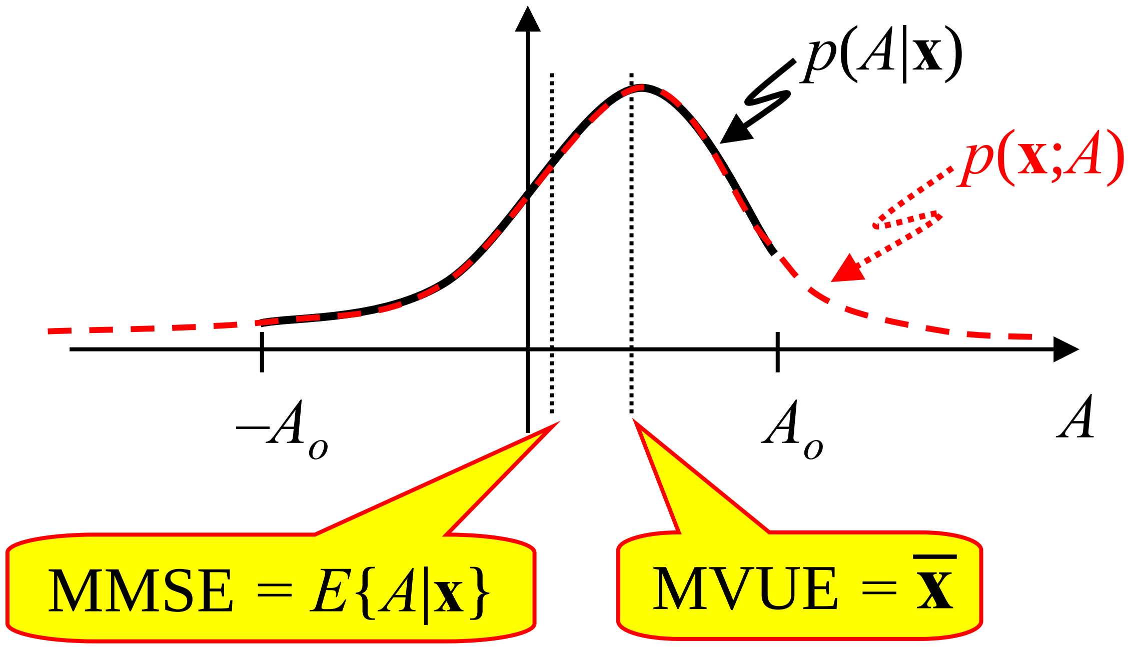
Conditional mean of A given data $\mathbf{x}$

$\underbrace{\text{Bayesian Minimum MSE}}_{\text{MMSE}}$ Estimate = The Mean of “posterior pdf”

So… we need to explore how to compute this from our data given knowledge of the Bayesian model for a problem

Compare this Bayesian Result to the Classical Result:

… for a given observed data vector x look at



<u>Before</u> taking any data… what is the best "estimate" of A?

- Classical: No best guess exists!
- Bayesian: Mean of the Prior PDF…
  - observed data "updates" this "*a priori*" estimate into an "*a posteriori*" estimate that balances "prior" vs. data

So… for this example we've seen that we need $E\{A|\mathbf{x}\}$.
How do we ***compute*** that!!!?? Well…

$$\hat{A} = E\{A \mid \mathbf{x}\}$$

$$= \int A p(A \mid \mathbf{x}) dA$$

So… we need the *posterior* pdf of A given the data… which can be found using Bayes' Rule:

Allows us to write one cond. PDF in terms of the other way around

$$p(A \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid A) p(A)}{p(\mathbf{x})}$$

$$= \frac{p(\mathbf{x} \mid A) p(A)}{\int p(\mathbf{x} \mid A) p(A) dA}$$

More easily found than $p(A|\mathbf{x})$… very much the same <u>structure</u> as the parameterized PDF used in Classical Methods

Assumed Known

So now we need $p(\mathbf{x}|A)$… For $x[n] = A + w[n]$ we know that

For A known, $x[n]$ is the known A plus random w[n]

$$p_x(x[n] \mid A) = p_w(x[n] - A \mid A)$$

Because $w[n]$ and A are assumed Independent

$$= p_w(x[n] - A)$$

PDF of x

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (x[n] - A)^2 \right]$$

Because $w[n]$ is White Gaussian they are independent… thus, the data conditioned on A is independent:

$$p(\mathbf{x} \mid A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right]$$

Same structure as the parameterized PDF used in Classical Methods…
But here A is an RV upon which we have conditioned the PDF!!!

Now we can use all this to find the MMSE for this problem:

Idea Easy!!

$$\hat{A} = E\{A \mid \mathbf{x}\} = \int A p(A \mid \mathbf{x}) dA = \frac{\int A \boxed{p(\mathbf{x} \mid A)} \boxed{p(A)} dA}{\int \boxed{p(\mathbf{x} \mid A)} \boxed{p(A)} dA}$$

Using Bayes' Rule

$$= \frac{\int_{-A_o}^{A_o} A \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - A\right)^2 \right] \left[1/2A_o\right] dA}{\int_{-A_o}^{A_o} \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - A\right)^2 \right] \left[1/2A_o\right] dA}$$

Use Prior PDF

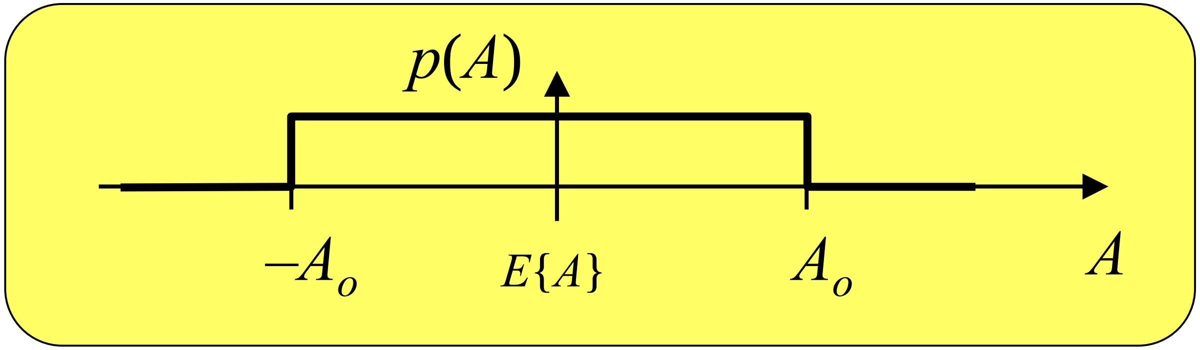
Use Parameter-Conditioned PDF

Hard to "Build"

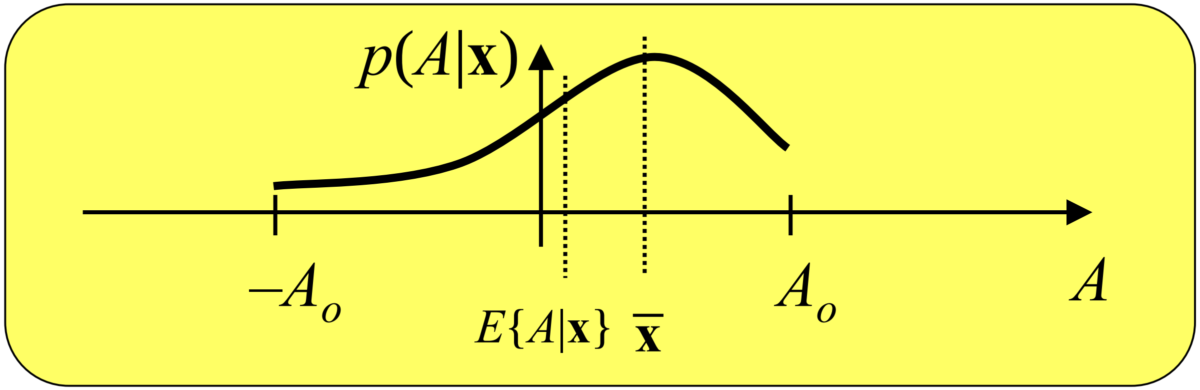
$$\hat{A} = \frac{\int_{-A_o}^{A_o} A \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - A\right)^2 \right] dA}{\int_{-A_o}^{A_o} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - A\right)^2 \right] dA}$$

<u>MMSE Estimator</u>… A function that maps observed data into the estimate… No Closed Form for this Case!!!

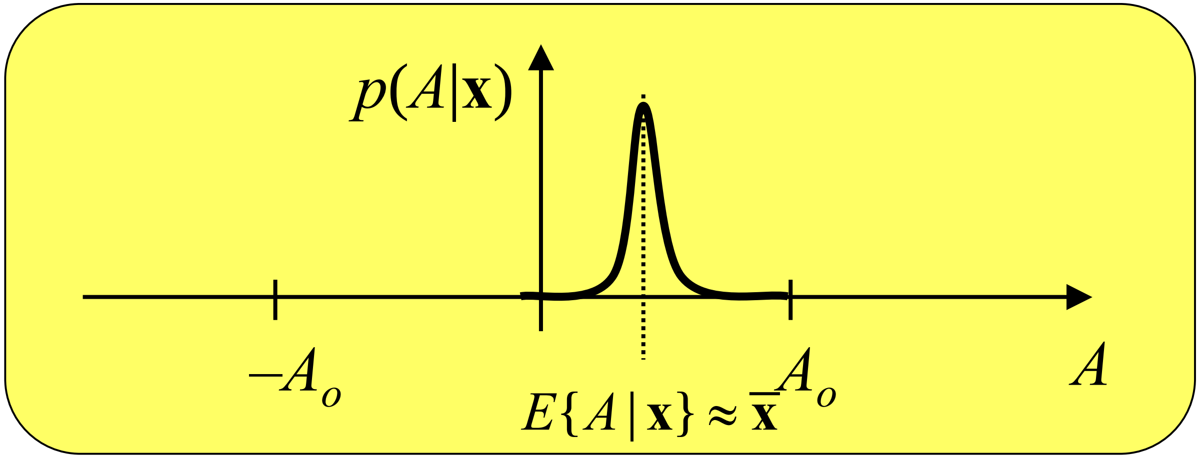
How the Bayesian approach balances *a priori* and *a posteriori* info:



No Data



Short Data Record



Long Data Record

General Insights From Example

1. After collecting data: our knowledge is captured by the posterior PDF $p(\theta|\mathbf{x})$

2. Estimator that minimizes the Bmse is $E\{\theta|\mathbf{x}\}$… the mean of the posterior PDF

3. Choice of prior is crucial:

   Bad Assumption of Prior $\Rightarrow$ Bad Bayesian Estimate!

   (Especially for short data records)

4. Bayesian MMSE estimator always <u>exists</u>!

   But <u>not necessarily</u> in <u>closed form</u>

   (Then must use numerical integration)

10.4 Choosing a Prior PDF

Choice is crucial:

1. Must be able to justify it physically

2. Anything other than a Gaussian prior will likely result in no closed-form estimates

We just saw that a uniform prior led to a non-closed form

We'll see here an example where a Gaussian prior gives a closed form

So… there seems to be a trade-off between:

- Choosing the prior PDF as accurately as possible
- Choosing the prior PDF to give computable closed form

Ex. 10.1: DC in WGN with Gaussian Prior PDF

We assume our Bayesian model is now: $x[n] = A + w[n]$

with a prior PDF of

$$A \sim N(\mu_A, \sigma_A^2)$$

AWGN

So… for a given value of the RV A the conditional PDF is

$$p(\mathbf{x} \mid A) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(x[n] - A\right)^2\right]$$

Then to get the needed conditional PDF we use this and the *a priori* PDF for A in Bayes' Theorem:

$$p(A \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid A)\,p(A)}{\int p(\mathbf{x} \mid A)\,p(A)\,dA}$$

Then… after much algebra and gnashing of teeth we get:

See the Book

$$p(A \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2_{A|x}}} \exp\left[-\frac{1}{2\sigma^2_{A|x}} \left(A - \mu_{A|x}\right)^2 \right]$$

which is a Gaussian PDF with

$$\mu_{A|x} = \left(\frac{N\sigma^2_{A|x}}{\sigma^2} \right) \bar{x} + \left(\frac{\sigma^2_{A|x}}{\sigma^2_A} \right) \mu_A$$

Weighted Combination of *a priori* and sample means

$$\sigma^2_{A|x} = \frac{1}{\dfrac{N}{\sigma^2} + \dfrac{1}{\sigma^2_A}}$$

“Parallel” Combination of *a priori* and sample variances

So… the main point here so far is that by assuming:

- Gaussian noise
- Gaussian *a priori* PDF on the parameter

We get a <u>Gaussian</u> *a posteriori* PDF for Bayesian estimation!!

Now recall that the Bayesian MMSE was the conditional *a posteriori* mean: $\boxed{\hat{A} = E\{A \mid \mathbf{x}\}}$

Because we now have a Gaussian *a posteriori* PDF it is easy to find an expression for this:

$$\hat{A} = E\{A \mid \mathbf{x}\} = \mu_{A|x} = \left(\frac{N\sigma^2_{A|x}}{\sigma^2}\right)\bar{x} + \left(\frac{\sigma^2_{A|x}}{\sigma^2_A}\right)\mu_A$$

$$\mathrm{var}\{\hat{A}\} = \mathrm{var}\{A \mid \mathbf{x}\} = \sigma^2_{A|x} = \frac{1}{\dfrac{N}{\sigma^2} + \dfrac{1}{\sigma^2_A}}$$

After some algebra we get:

$$\boxed{\begin{aligned}\hat{A} &= \left(\frac{\sigma^2_A}{\sigma^2_A + \dfrac{\sigma^2}{N}}\right)\bar{x} + \left(\frac{\dfrac{\sigma^2}{N}}{\sigma^2_A + \dfrac{\sigma^2}{N}}\right)\mu_A \\ &= \alpha\bar{x} + (1-\alpha)\mu_A, \qquad 0 < \alpha < 1\end{aligned}}$$

Easily Computable Estimator:

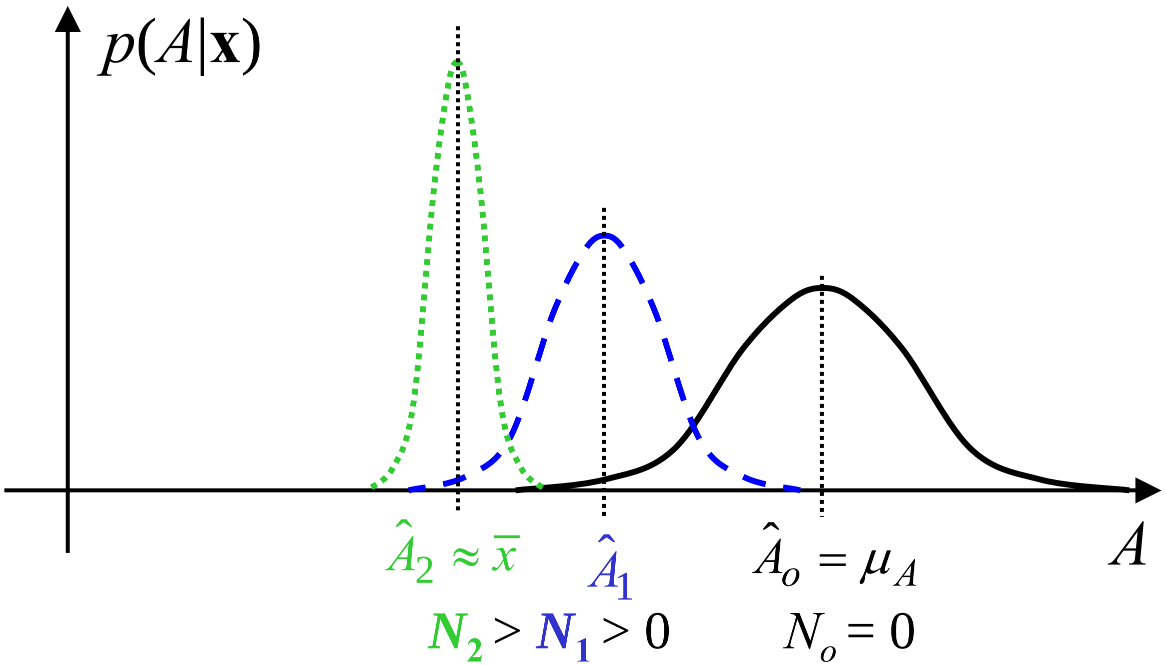
- Sample mean computed from data
- σ known from data model
- μ_A and σ_A known from prior model

Little or Poor Data: $\sigma^2_A << \sigma^2/N \qquad \hat{A} \approx \mu_A$

Much or Good Data: $\sigma^2_A >> \sigma^2/N \qquad \hat{A} \approx \bar{x}$

Comments on this Example for Gaussian Noise and Gaussian Prior

1. Closed-Form Solution for Estimate!
2. Estimate is… Weighted sum of prior mean & data mean
3. Weights balance between prior info quality and data quality
4. As N increases…
   a. Estimate $E\{A|\mathbf{x}\}$ moves $\mu_A \rightarrow \bar{x}$
   b. Accuracy var$\{A|\mathbf{x}\}$ moves $\sigma_A^2 \rightarrow \sigma^2/N$



Bmse for *this* Example: $Bmse(\hat{A}) = \sigma^2_{A|x}$

To see this:

$$Bmse\left(\hat{A}\right) = E\left\{\left(A - \hat{A}\right)^2\right\}$$

$$= \iint \left(A - \hat{A}\right)^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA$$

$$= \iint \left(A - E\{A \mid \mathbf{x}\}\right)^2 p\left(A|\mathbf{x}\right) p(\mathbf{x})\, d\mathbf{x}\, dA$$

$$= \int \underbrace{\left[\int \left(A - E\{A \mid \mathbf{x}\}\right)^2 p\left(A|\mathbf{x}\right) dA\right]}_{=\mathrm{var}\{A|\mathbf{x}\} = \sigma^2_{A|x}} p(\mathbf{x})\, d\mathbf{x}$$

General Result: Bmse = posterior variance averaged over PDF of x

In this case $\sigma_{A|x}$ is not a function of $\mathbf{x}$:

$$Bmse\left(\hat{A}\right) = \sigma^2_{A|x} \int p(\mathbf{x})\, d\mathbf{x} = \sigma^2_{A|x}$$

The big thing that this example shows:

Gaussian Data & Gaussian Prior gives Closed-Form MMSE Solution

This will hold in general!