FSAN/ELEG815: Statistical Learning

Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware

6. Maximum Likelihood and Bayes Estimation

Definition (Bayes Estimation)

Objective: Estimate a random parameter (*RV*) from observations samples $x_1, x_2, \cdots, x_n$ that are statistically related to y by $f_{y|\mathbf{x}}(\cdot)$

Bayes Procedure: Define a nonnegative cost function $C(y, \hat{y})$ and set $\hat{y}$ to minimize the expected cost, or risk

$$\underbrace{R}_{\text{risk}} = E\{C(y, \hat{y})\}$$

Since y and $\hat{y}$ are *RVs*

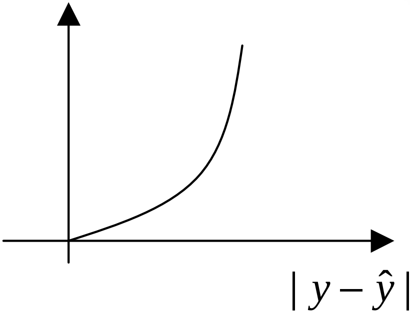
$$\begin{aligned} R &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(y, \hat{y}) f_{y,\mathbf{x}}(y, \mathbf{x}) dy d\mathbf{x} \\ &= \int_{-\infty}^{\infty} \underbrace{\left[\int_{-\infty}^{\infty} C(y, \hat{y}) f_{y|\mathbf{x}}(y|\mathbf{x}) dy\right]}_{I(\hat{y})} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Note: Minimizing $I(\hat{y})$ it is equivalent to minimize R since $f_{\mathbf{x}}(\mathbf{x}) \geq 0$

Consider several cost functions

Case 1: Mean Squared cost function

$$C(y,\hat{y}) = |y-\hat{y}|^2$$



In this case,

$$\begin{aligned} I(\hat{y}) &= \int_{-\infty}^{\infty} (y-\hat{y})^2 f_{y|\mathbf{x}}(y|\mathbf{x})dy \\ \Rightarrow \frac{\partial I(\hat{y})}{\partial \hat{y}} &= -2\int_{-\infty}^{\infty} (y-\hat{y}) f_{y|\mathbf{x}}(y|\mathbf{x})dy = 0 \end{aligned}$$

or rearranging

$$\int_{-\infty}^{\infty} \hat{y} f_{y|\mathbf{x}}(y|\mathbf{x})dy = \int_{-\infty}^{\infty} y f_{y|\mathbf{x}}(y|\mathbf{x})dy \qquad [\hat{y} \text{ is a constant}]$$

$$\Rightarrow \hat{y}_{\text{MS}} = \int_{-\infty}^{\infty} y f_{y|\mathbf{x}}(y|\mathbf{x})dy = E\{y|\mathbf{x}\}$$

Example

Let $x_i = a + \mu_i$ for $i = 1, 2, \cdots, N$, where $\mu_i \sim N(0, \sigma^2)$ and $a \sim N(0, \sigma_a^2)$ are i.i.d. Determine $\hat{a}_{\text{MS}}(\mathbf{x})$.

Note

$$f_{\mathbf{x}|a}(\mathbf{x}|a) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-a)^2}{2\sigma^2}} = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2}\left(\sum_{i=1}^{N} \frac{(x_i-a)^2}{\sigma^2}\right)} \qquad (*)$$

$$f_a(a) = \frac{1}{\sqrt{2\pi}\sigma_a} e^{-\frac{a^2}{2\sigma_a^2}} \qquad (**)$$

To find $\hat{a}_{\text{MS}}(\mathbf{x})$ we need

$$\hat{a}_{\text{MS}}(\mathbf{x}) = E\{a|\mathbf{x}\}$$

By Bayes's theorem we can write

$$f_{a|\mathbf{x}}(a|\mathbf{x}) = \frac{f_{\mathbf{x}|a}(\mathbf{x}|a)f_a(a)}{f_{\mathbf{x}}(\mathbf{x})}$$

Substituting in $(*)$ and $(**)$, and rearranging

$$f_{a|\mathbf{x}}(a|\mathbf{x}) = \frac{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}}\left(\frac{1}{\sqrt{2\pi}\sigma_a}\right)e^{-\frac{1}{2}\left(\sum_{i=1}^{N}\frac{(x_i-a)^2}{\sigma^2}+\frac{a^2}{\sigma_a^2}\right)}}{f_{\mathbf{x}}(\mathbf{x})}$$

This can be compactly written as

$$f_{a|\mathbf{x}}(a|\mathbf{x}) = C(\mathbf{x})\exp\left\{-\frac{1}{2\sigma_p^2}\left[a-\frac{\sigma_a^2}{\sigma_a^2+\sigma^2/N}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)\right]^2\right\}$$

Observations on

$$
\begin{aligned}
f_{a|\mathbf{x}}(a|\mathbf{x}) &= C(\mathbf{x}) \exp\left\{ -\frac{1}{2\sigma_p^2} \left[a - \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2/N} \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right) \right]^2 \right\} \\
&= C(\mathbf{x}) \exp\left\{ -\frac{(a-\eta)^2}{2\sigma_p^2} \right\}
\end{aligned}
$$

- $C(\mathbf{x})$ is a (normalizing) function of $\mathbf{x}$ only
- The variance term is given by

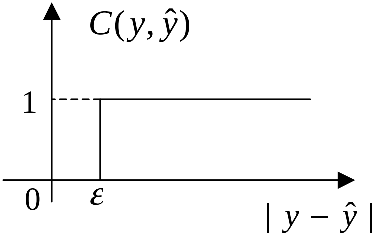
$$
\sigma_p^2 = \left(\frac{1}{\sigma_a^2} + \frac{N}{\sigma^2} \right)^{-1} = \frac{\sigma_a^2 \sigma^2}{N\sigma_a^2 + \sigma^2}
$$

- Critical Observation: $f_{a|\mathbf{x}}(a|\mathbf{x})$ is a Gaussian distribution!
- Result:

$$
\hat{a}_{\text{MS}} = E\{a|\mathbf{x}\} = \eta = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2/N} \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right)
$$

Case 2: Uniform cost function

$$C(y,\hat{y}) = \begin{cases} 0 & |y-\hat{y}| < \epsilon \\ 1 & \text{else} \end{cases}$$



Question: For what types of problems is this cost function effective?

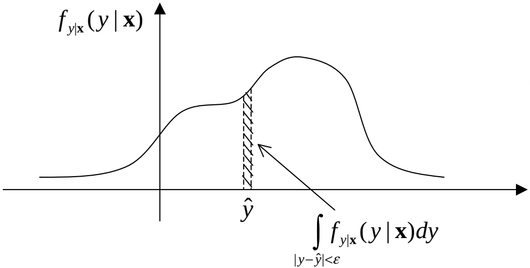
In this case,

$$\begin{aligned} I(\hat{y}) &= \int_{-\infty}^{\infty} C(y,\hat{y}) f_{y|\mathbf{x}}(y|\mathbf{x})dy \\ &= \int_{|y-\hat{y}|\geq\epsilon} f_{y|\mathbf{x}}(y|\mathbf{x})dy \\ &= 1 - \int_{|y-\hat{y}|<\epsilon} f_{y|\mathbf{x}}(y|\mathbf{x})dy \end{aligned}$$

How do we minimize $I(\hat{y})$?

Result: $I(\hat{y})$ is minimized by maximizing

$$\int_{|y-\hat{y}|<\epsilon} f_{y|\mathbf{x}}(y|\mathbf{x})dy$$



Note: ϵ is arbitrarily small

$\Rightarrow$ $I(\hat{y})$ is minimized when $f_{y|\mathbf{x}}(y|\mathbf{x})$ takes its largest value

$$\hat{y}_{\text{MAP}}(\mathbf{x}) = \underset{y}{\operatorname{argmax}}\, f_{y|\mathbf{x}}(y|\mathbf{x})$$

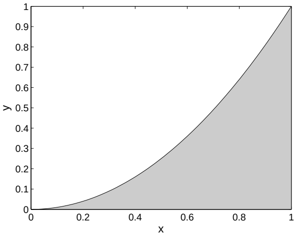
- $\hat{y}_{\text{MAP}}$ is referred to as the maximum a posteriori (MAP) estimate because it maximized the posterior density $f_{y|\mathbf{x}}(y|\mathbf{x})$.

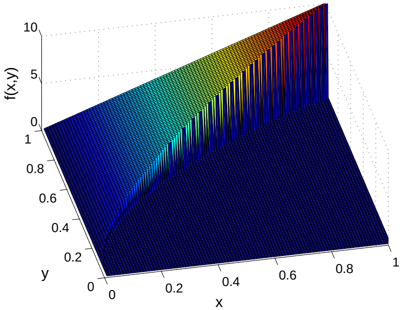
Example

Let

$$f_{x,y}(x,y) = \begin{cases} 10y & 0 \leq y \leq x^2, 0 \leq x \leq 1 \\ 0 & \text{otherwie} \end{cases}$$

Find the MS and MAP estimates of y, i.e., $\hat{y}_{\text{MS}}(x)$ and $\hat{y}_{\text{MAP}}(x)$.





Example

Let

$$f_{x,y}(x,y) = \begin{cases} 10y & 0 \le y \le x^2, 0 \le x \le 1 \\ 0 & \text{otherwie} \end{cases}$$

Find the MS and MAP estimates of y, i.e., $\hat{y}_{\text{MS}}(x)$ and $\hat{y}_{\text{MAP}}(x)$.

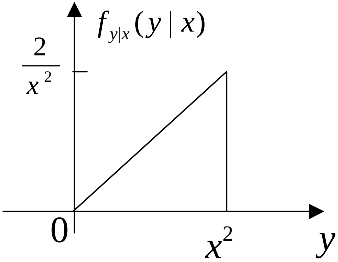
First step: determine the posterior density $f_{y|x}(y|x)$.

Since $f_{y|x}(y|x) = \frac{f_{x,y}(x,y)}{f_x(x)}$, we need

$$\begin{aligned} f_x(x) &= \int_{-\infty}^{\infty} f_{x,y}(x,y)dy \\ &= \int_0^{x^2} 10y dy \\ &= 5y^2 \Big|_0^{x^2} = 5x^4 \quad 0 \le x \le 1 \end{aligned}$$

Thus,

$$
\begin{aligned}
f_{y|x}(y|x) &= \frac{f_{x,y}(x,y)}{f_x(x)} \\
&= \frac{10y}{5x^4} = \frac{2y}{x^4} \qquad 0 \le y \le x^2
\end{aligned}
$$



MAP estimate:

$$
\begin{aligned}
\hat{y}_{\mathsf{MAP}}(x) &= \underset{y}{\operatorname{argmax}}\, f_{y|x}(y|x) \\
&= \underset{y}{\operatorname{argmax}}\, \frac{2y}{x^4} \quad 0 \le y \le x^2 \\
&= x^2
\end{aligned}
$$

MS estimate:

$$
\begin{aligned}
\hat{y}_{\mathsf{MS}}(x) &= E\{y|x\} \\
&= \int_0^{x^2} y f_{y|x}(y|x)dy \\
&= \int_0^{x^2} \frac{2y^2}{x^4} dy \\
&= \frac{2}{3}\frac{y^3}{x^4}\Big|_0^{x^2} = \frac{2}{3}x^2
\end{aligned}
$$

Note that the minimum MSE is

$$\begin{aligned} E\{(y-\hat{y}_{\mathsf{MS}})^2\} &= \int_0^1\int_0^{x^2}(y-\hat{y}_{\mathsf{MS}})^2 f_{x,y}(x,y)dydx \\ &= \int_0^1\int_0^{x^2}(y-\frac{2}{3}x^2)^2 10ydydx = \frac{5}{162} = 0.0309 \end{aligned}$$

The MSE of the MAP estimate is

$$\begin{aligned} E\{(y-\hat{y}_{\mathsf{MAP}})^2\} &= \int_0^1\int_0^{x^2}(y-\hat{y}_{\mathsf{MAP}})^2 f_{x,y}(x,y)dydx \\ &= \int_0^1\int_0^{x^2}(y-x^2)^2 10ydydx = \frac{5}{54} = 0.0926 \end{aligned}$$

Observation: This result is expected. Why?

Observation: MAP estimation can be used as an extension of ML estimation if some variability is assumed

- Instead of an unknown constant θ, we have an unknown random parameter with distribution $f_\theta(\theta)$

To see this, note

$$f_{\theta|\mathbf{x}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)f_\theta(\theta)}{f_{\mathbf{x}}(\mathbf{x})}$$

- The MAP estimate maximizes the numerator since $f_{\mathbf{x}}(\mathbf{x})$ is not a function of θ,

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\, f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)f_\theta(\theta)$$

Question: For what distribution $f_\theta(\theta)$ does $\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{ML}}$?
That is

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\, f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)f_\theta(\theta) \stackrel{?}{=} \underset{\theta}{\operatorname{argmax}}\, f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) = \hat{\theta}_{\text{ML}}$$

Example

Let $x(n) = A + \mu(n)$ for $n = 1, 2, \cdots, N$, where $\mu(n) \sim N(0, \sigma_\mu^2)$ and $A \sim N(A_0, \sigma_A^2)$ are i.i.d.

Determine the MAP estimate of A.

Need to maximize $f_{\mathbf{x}|A}(\mathbf{x}|A) f_A(A)$, or

$$\hat{A}_{\text{MAP}} = \underset{A}{\operatorname{argmax}} \left[\ln(f_{\mathbf{x}|A}(\mathbf{x}|A)) + \ln(f_A(A)) \right]$$

Note

$$\ln(f_{\mathbf{x}|A}(\mathbf{x}|A)) = \frac{N}{2} \ln\left(\frac{1}{2\pi\sigma_\mu^2}\right) - \sum_{n=0}^{N} \frac{(x(n) - A)^2}{2\sigma_\mu^2}$$

and

$$\ln(f_A(A)) = \frac{1}{2} \ln\left(\frac{1}{2\pi\sigma_A^2}\right) - \frac{(A - A_0)^2}{2\sigma_A^2}$$

Thus

$$\hat{A}_{\text{MAP}} = \underset{A}{\operatorname{argmin}} \left(\frac{1}{2\sigma_\mu^2} \sum_{n=1}^{N} (x(n) - A)^2 + \frac{(A - A_0)^2}{2\sigma_A^2} \right)$$

Differentiating we get

$$-\frac{1}{\sigma_\mu^2} \sum_{n=1}^{N} (x(n) - A) + \frac{(A - A_0)}{\sigma_A^2} \Bigg|_{A=\hat{A}_{\text{MAP}}} = 0$$

$$\Rightarrow \sum_{n=1}^{N} \frac{x(n)}{\sigma_\mu^2} - \frac{N\hat{A}_{\text{MAP}}}{\sigma_\mu^2} = \frac{\hat{A}_{\text{MAP}}}{\sigma_A^2} - \frac{A_0}{\sigma_A^2}$$

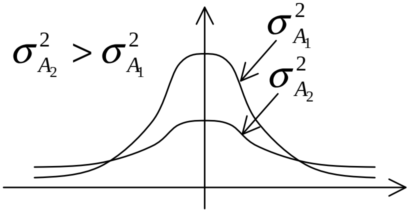
$$\Rightarrow \hat{A}_{\text{MAP}} \left(\frac{1}{\sigma_A^2} + \frac{N}{\sigma_\mu^2} \right) = \frac{1}{\sigma_\mu^2} \sum_{n=1}^{N} x(n) + \frac{A_0}{\sigma_A^2}$$

$$\Rightarrow \hat{A}_{\text{MAP}} = \frac{1}{\frac{1}{\sigma_A^2} + \frac{N}{\sigma_\mu^2}} \left(\frac{1}{\sigma_\mu^2} \sum_{n=1}^{N} x(n) + \frac{A_0}{\sigma_A^2} \right)$$

$$\hat{A}_{\text{MAP}} = \frac{1}{\frac{1}{\sigma_A^2} + \frac{N}{\sigma_\mu^2}} \left(\frac{1}{\sigma_\mu^2} \sum_{n=1}^{N} x(n) + \frac{A_0}{\sigma_A^2} \right)$$

Note that if $\sigma_A^2 \rightarrow \infty$ then there is no *a priori* information and

$$\lim_{\sigma_A^2 \rightarrow \infty} \hat{A}_{\text{MAP}} = \frac{1}{N} \sum_{n=1}^{N} x(n) = \hat{A}_{\text{ML}}$$

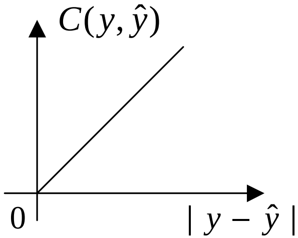


Observation: As $f_\theta(\theta)$ flattens out

$$\hat{\theta}_{\text{MAP}} \rightarrow \hat{\theta}_{\text{ML}}$$

Case 3: The absolute cost function

$$C(y,\hat{y}) = |y-\hat{y}|$$



Question: For what types of problems is this cost function effective?

In this case

$$\begin{aligned} I(\hat{y}) &= \int_{-\infty}^{\infty} C(y,\hat{y}) f_{y|\mathbf{x}}(y|\mathbf{x})dy \\ &= \int_{y<\hat{y}} (\hat{y}-y) f_{y|\mathbf{x}}(y|\mathbf{x})dy + \int_{y\geq\hat{y}} (y-\hat{y}) f_{y|\mathbf{x}}(y|\mathbf{x})dy \\ &= \int_{-\infty}^{\hat{y}} (\hat{y}-y) f_{y|\mathbf{x}}(y|\mathbf{x})dy + \int_{\hat{y}}^{\infty} (y-\hat{y}) f_{y|\mathbf{x}}(y|\mathbf{x})dy \end{aligned}$$

$$I(\hat{y}) = \int_{-\infty}^{\hat{y}} (\hat{y} - y) f_{y|\mathbf{x}}(y|\mathbf{x})dy + \int_{\hat{y}}^{\infty} (y - \hat{y}) f_{y|\mathbf{x}}(y|\mathbf{x})dy$$

Note that

$$\int_{-\infty}^{\hat{y}} (\hat{y} - y) f_{y|\mathbf{x}}(y|\mathbf{x})dy = \hat{y} F_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) - \int_{-\infty}^{\hat{y}} y f_{y|\mathbf{x}}(y|\mathbf{x})dy$$

and similarly

$$\int_{\hat{y}}^{\infty} (y - \hat{y}) f_{y|\mathbf{x}}(y|\mathbf{x})dy = \int_{\hat{y}}^{\infty} y f_{y|\mathbf{x}}(y|\mathbf{x})dy - \hat{y}(1 - F_{y|\mathbf{x}}(\hat{y}|\mathbf{x}))$$

Thus,

$$\begin{aligned} I(\hat{y}) &= \hat{y} F_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) - \int_{-\infty}^{\hat{y}} y f_{y|\mathbf{x}}(y|\mathbf{x})dy \\ &\quad - \hat{y}(1 - F_{y|\mathbf{x}}(\hat{y}|\mathbf{x})) + \int_{\hat{y}}^{\infty} y f_{y|\mathbf{x}}(y|\mathbf{x})dy \end{aligned}$$

$$
\begin{aligned}
I(\hat{y}) &= \hat{y}F_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) - \int_{-\infty}^{\hat{y}} y f_{y|\mathbf{x}}(y|\mathbf{x})dy \\
&\quad -\hat{y}(1 - F_{y|\mathbf{x}}(\hat{y}|\mathbf{x})) + \int_{\hat{y}}^{\infty} y f_{y|\mathbf{x}}(y|\mathbf{x})dy
\end{aligned}
$$

Taking the derivative

$$
\begin{aligned}
\frac{\partial I(\hat{y})}{\partial \hat{y}} &= F_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) + \hat{y}f_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) - \hat{y}f_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) \\
&\quad -(1 - F_{y|\mathbf{x}}(\hat{y}|\mathbf{x})) + \hat{y}f_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) - \hat{y}f_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) \\
&= F_{y|\mathbf{x}}(\hat{y}|\mathbf{x}) - (1 - F_{y|\mathbf{x}}(\hat{y}|\mathbf{x}))
\end{aligned}
$$

Result: Setting equal to 0, we see the $\hat{y}_{\text{MAE}}$ is given by

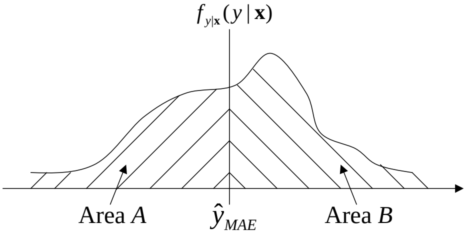
$$
F_{y|\mathbf{x}}(\hat{y}_{\text{MAE}}|\mathbf{x}) = 1 - F_{y|\mathbf{x}}(\hat{y}_{\text{MAE}}|\mathbf{x})
$$

or

$$
\int_{-\infty}^{\hat{y}_{\text{MAE}}} f_{y|\mathbf{x}}(y|\mathbf{x})dy = \int_{\hat{y}_{\text{MAE}}}^{\infty} f_{y|\mathbf{x}}(y|\mathbf{x})dy
$$

$$\int_{-\infty}^{\hat{y}_{\text{MAE}}} f_{y|\mathbf{x}}(y|\mathbf{x})dy = \int_{\hat{y}_{\text{MAE}}}^{\infty} f_{y|\mathbf{x}}(y|\mathbf{x})dy$$

Interpreting this graphically



$$\text{Area } A = \text{Area } B$$

Observation:

$$\hat{y}_{\text{MAE}} = \text{median of } f_{y|\mathbf{x}}(y|\mathbf{x})$$

Estimator Relations

- If $f_{y|\mathbf{x}}(y|\mathbf{x})$ is symmetric, then

$$\hat{y}_{\text{MAE}} = \hat{y}_{\text{MS}}$$

  Why?
  For a symmetric distribution the conditional mean is equal to the (median) symmetry point

- If $f_{y|\mathbf{x}}(y|\mathbf{x})$ is symmetric and unimodal, then

$$\hat{y}_{\text{MAE}} = \hat{y}_{\text{MS}} = \hat{y}_{\text{MAP}}$$

  Why? The unimodal constraint implies that the single mode must be at the distribution symmetry point $\Rightarrow$ the MAP estimate is located at the central point

Example

Determine $\hat{y}_{\text{MAE}}$ for the previously considered case

$$f_{x,y}(x,y) = \begin{cases} 10y & 0 \le y \le x^2, 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

We showed previously that

$$f_{y|x}(y|x) = \frac{2y}{x^4} \quad 0 \le y \le x^2 \qquad \Rightarrow \qquad F_{y|x}(y|x) = \frac{y^2}{x^4} \quad 0 \le y \le x^2$$

Thus determining $\hat{y}_{\text{MAE}}$

$$\begin{aligned} F_{y|x}(\hat{y}_{\text{MAE}}|x) &= 1 - F_{y|x}(\hat{y}_{\text{MAE}}|x) \\ \Rightarrow \frac{\hat{y}^2_{\text{MAE}}}{x^4} &= 1 - \frac{\hat{y}^2_{\text{MAE}}}{x^4} \\ \Rightarrow \hat{y}_{\text{MAE}} &= \frac{x^2}{\sqrt{2}} \end{aligned}$$

MAP estimate: (previous result)

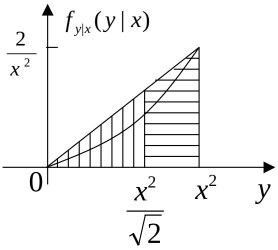
$$\hat{y}_{\text{MAP}}(x) = \underset{y}{\operatorname{argmax}}\, f_{y|x}(y|x) = x^2$$

MS estimate: (previous result)

$$\hat{y}_{\text{MS}}(x) = E\{y|x\} = \frac{2}{3}x^2$$

MAE estimate:

$$\hat{y}_{\text{MAE}}(x) = \text{median of } f_{y|\mathbf{x}}(y|\mathbf{x}) = \frac{x^2}{\sqrt{2}}$$



Final ML and MAP Comments

- ML estimation was pioneered by geneticist and statistician Sir R. A. Fisher between 1912 and 1922
- Under fairly weak regularity conditions the ML estimate is asymptotically optimal
  - The ML estimate is asymptotically unbiased, i.e., its bias tends to zero as the number of samples increases to infinity
  - The ML estimate is asymptotically efficient, i.e., it achieves the Cramér-Rao lower bound when the number of samples tends to infinity
    Consequence: No unbiased estimator has lower mean squared error than the ML estimator
  - The ML estimate is asymptotically normal, i.e., as the number of samples increases, the distribution of the ML estimate tends to the Gaussian distribution
- MAP estimation is a generalization of ML estimation that incorporates the prior distribution of the quantity being estimated