6.s077 — INTRODUCTION TO DATA SCIENCE
EECS, MIT, Spring 2018

Lecture 6

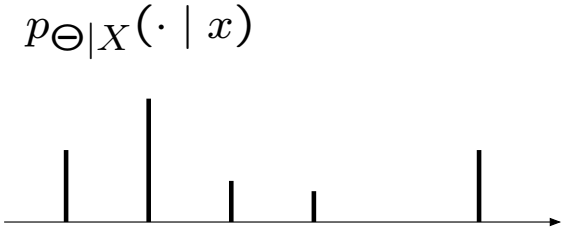
Bayesian Inference and an Example

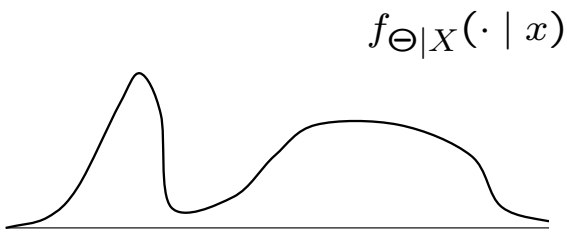
Today's agenda

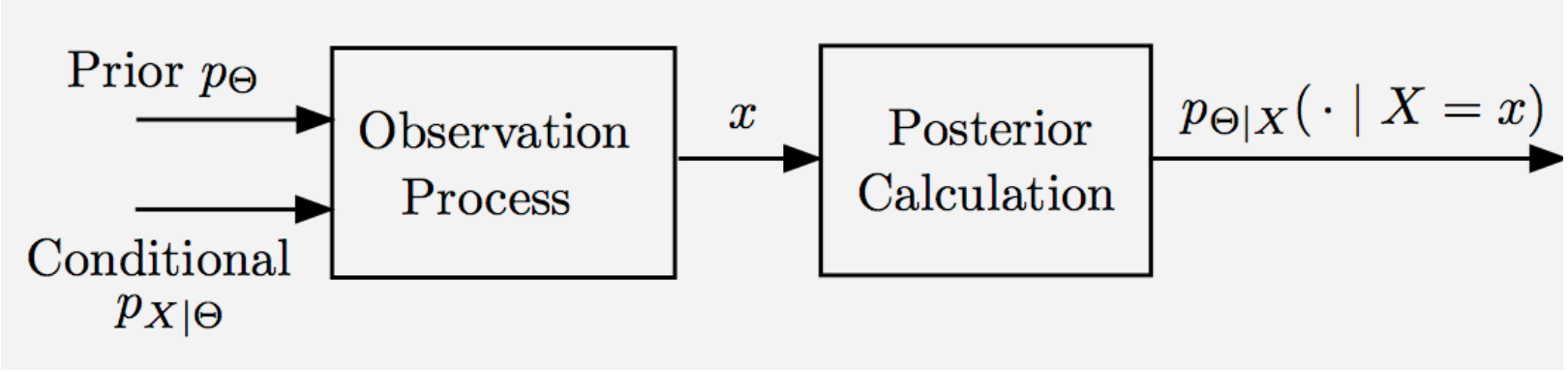
- Review of the Bayesian framework
  - the posterior distribution
  - Bayesian confidence intervals
- Point estimates
  - Maximum A Posteriori probability (MAP) estimator
  - conditional expectation (LMS estimator)
  - error analysis
- Bottlenecks in Bayesian inference
- Linear Least Mean Squares (LLMS estimation)
  - the one-dimensional case
  - generalizations
- Linear normal models
  - review of the normal distribution
  - tracking a moving vehicle
  - modeling and solution

Review of the Bayesian framework

$\mathbb{P}_\Theta$	$\mathbb{P}_{X\mid\Theta}$
Θ?	X







- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X = x)$ or $f_{\Theta|X}(\cdot \mid X = x)$

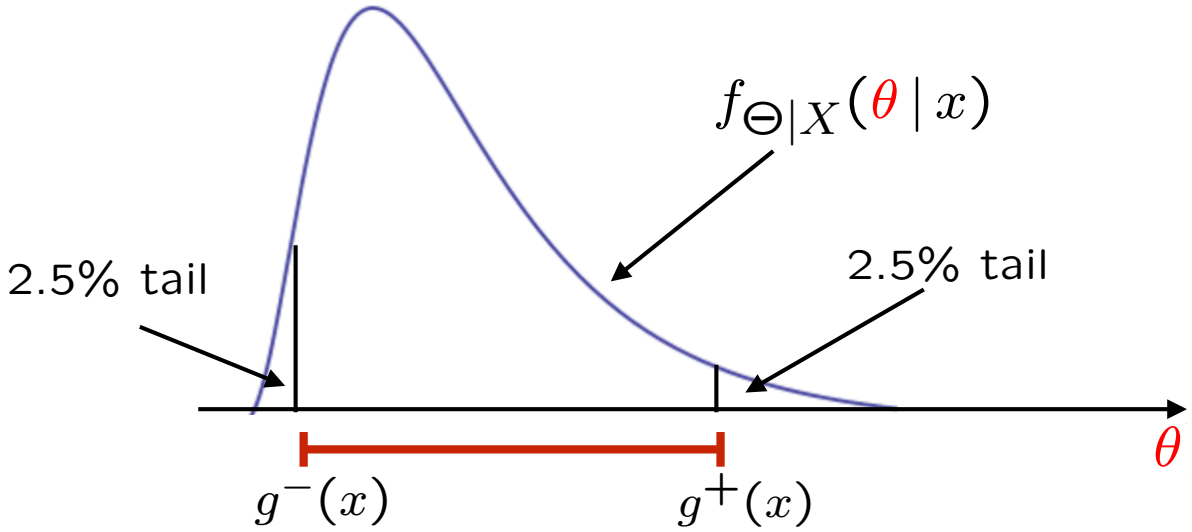
$$\mathbb{P}_{\Theta|X} = \frac{\mathbb{P}_\Theta \cdot \mathbb{P}_{X|\Theta}}{\mathbb{P}_X}$$

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta) \cdot p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta) \cdot f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

etc.

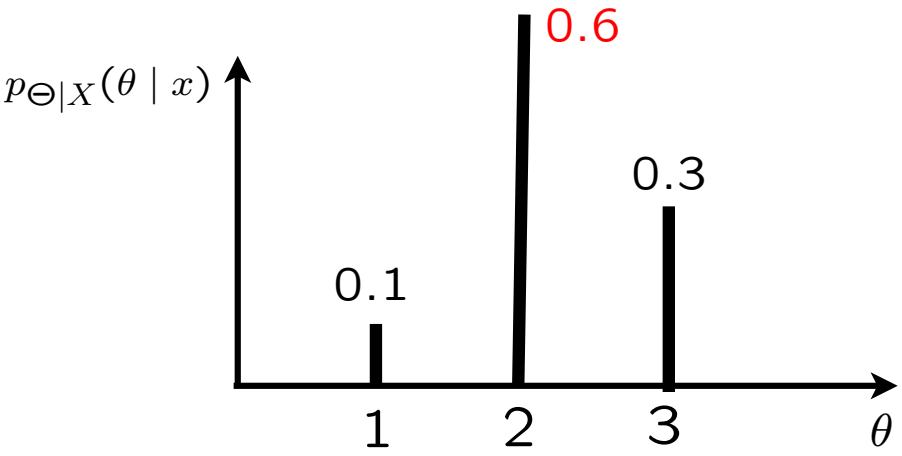
Bayesian confidence intervals



95% coinfidence interval

$$\mathbb{P}\Big(g^-(x) \leq \Theta \leq g^+(x) \,|\, X = x\Big) = 0.95$$

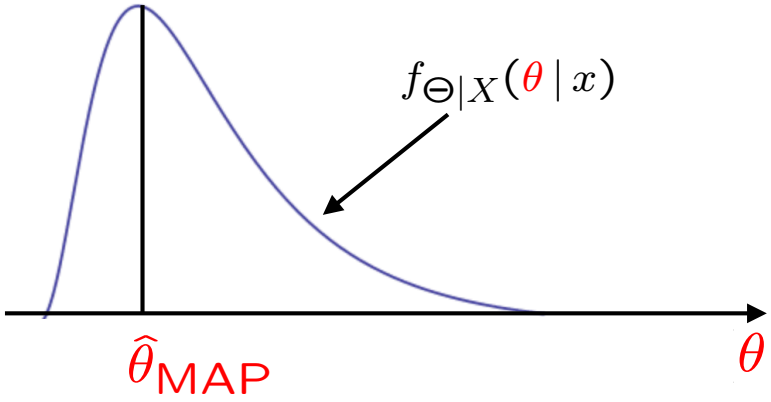
Point estimates - Maximum a Posteriori probability (MAP)

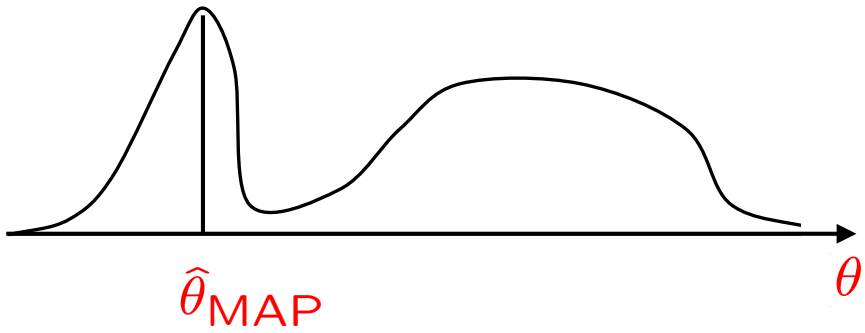


- conditional prob of error:

  $\mathbb{P}(\hat{\theta} \neq \Theta \mid X = x)$

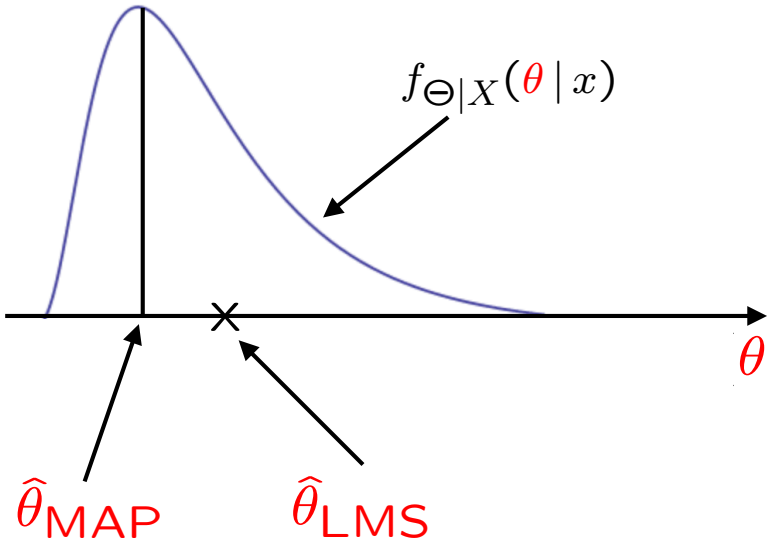
  smallest under the MAP rule

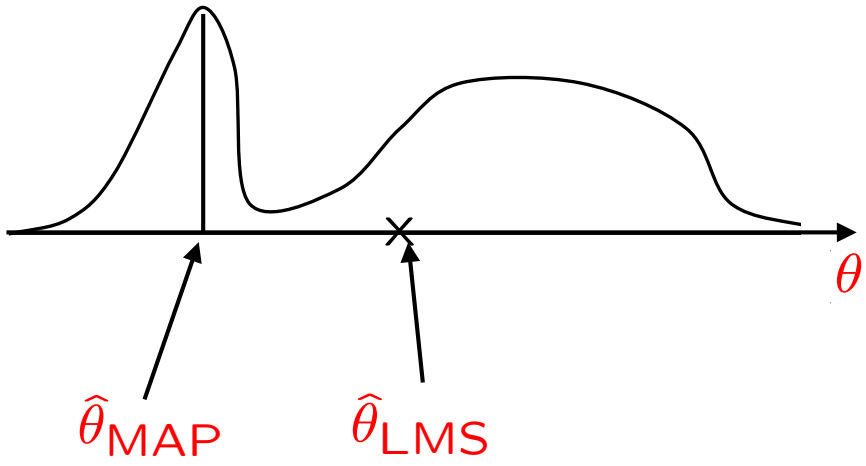




Point estimates - Conditional Expectation (LMS)

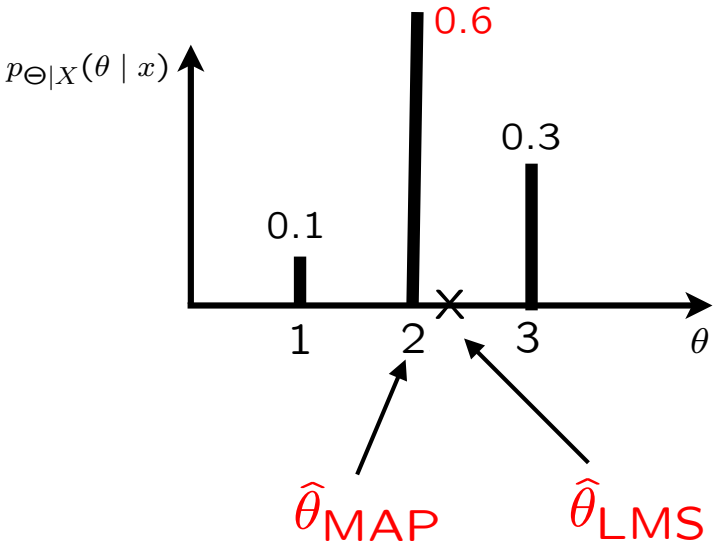
- $\widehat{\theta} = \mathbb{E}[\Theta \mid X = x]$





- Least Mean Squares property: minimizes $\mathbb{E}\Big[(\Theta - \widehat{\theta})^2 \mid X = x\Big]$
- Optimal conditional mean squared error is:
  $\mathbb{E}\Big[\Big(\Theta - \mathbb{E}[\Theta \mid X = x]\Big)^2 \mid X = x\Big]$: conditional variance
  (variance of the posterior distribution)
- $\mathbb{E}[\Theta \mid X]$ minimizes $\mathbb{E}\Big[(\Theta - \widehat{\Theta})^2\Big]$ over all estimators
- Most appropriate when θ is continuous

Comments on the LMS estimator



- LMS may not be appropriate when Θ is discrete/categorical

- Error analysis

$$\mathbb{E}\Big[(\Theta - \widehat{\Theta})^2\Big] = \mathbb{E}\Big[\mathbb{E}\Big[(\Theta - \widehat{\Theta})^2 \;\Big|\; X = x\Big]\Big]$$

(variance of the posterior distribution)

$$\int f_X(x)\Big[\int (\Theta - \mathbb{E}[\Theta \mid X = x])^2 f_{\Theta|X}(\theta \mid x)\, d\theta\Big]\, dx$$

$$\int\int f_{\Theta,X}(\theta, x)\Big(\Theta - \mathbb{E}[\Theta \mid X = x]\Big)^2\, d\theta\, dx$$

can be messy

The Bayesian LMS bottlenecks

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x \mid \theta')\, d\theta'$$

- Full correct model, $f_{X|\Theta}(x \,|\, \theta)$, may not be available
  - may need previous stage of "statistics" to build the model
- Can be hard to compute/implement/analyze
  - vector $\theta \longrightarrow$ multidimensional integral in denominator
  - mean squared error calculation is even worse
  - (MAP is easier, can ignore the denominator)
- It may help to consider restricted classes of estimators

Linear Least Mean Squares (LLMS) estimation

- Have probabilistic model of (Θ, X)
- Restrict to estimators of the form $\widehat{\Theta} = aX + b$
  - choose a, b to minimize $\mathbb{E}\Big[(\Theta - aX - b)^2\Big]$
  - quadratic in a, b
  - set derivatives (w.r.t.a, b) to zero, do algebra...

$$\widehat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)}\big(X - \mathbf{E}[X]\big)$$

- Only means, variances, covariances matter

$$\mathbf{E}\Big[(\widehat{\Theta}_L - \Theta)^2\Big] = (1 - \rho^2)\,\mathrm{var}(\Theta)$$

Linear Least Mean Squares (LLMS) — useful facts

- Generalizes to the case of vector Θ and X: $\widehat{\Theta} = LX + b$
- Find matrix L, vector b that minimize $\mathbb{E}\Big[\|\Theta - LX - b\|^2\Big]$

$$\widehat{\Theta}_L = \mathbb{E}[\Theta] + \Big[\text{Cov}(\Theta_i, X_j)\Big]\Big[\text{Cov}(X_j, X_k)\Big]^{-1}\Big(X - \mathbb{E}[X]\Big)$$

matrices

- Linear normal models: all linear functions of (X, Θ), are normal
  - turns out that $\widehat{\Theta}_{LMS}$ is linear, given by same formula as for $\widehat{\Theta}_L$
- LLMS estimation is equivalent to assuming/pretending that we have a linear normal model
- Choice of what we call "data" (X) can make a difference
  $\widehat{\Theta} = aX + b$ versus $\widehat{\Theta} = a \log X + b$
  think of $\log X$ as playing the role of X

Recognizing normal PDFs

$X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$c \cdot e^{-8(x-3)^2}$ μ? σ?

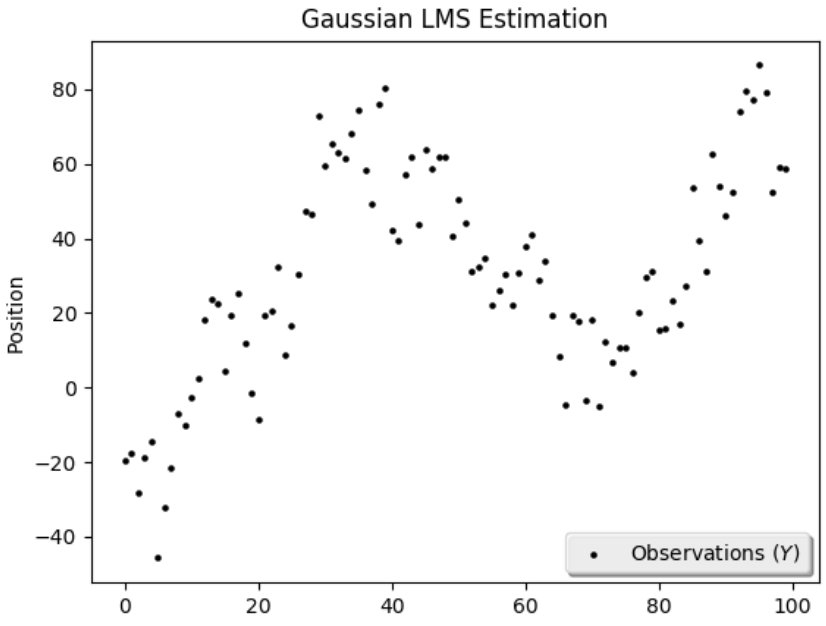
$$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$$

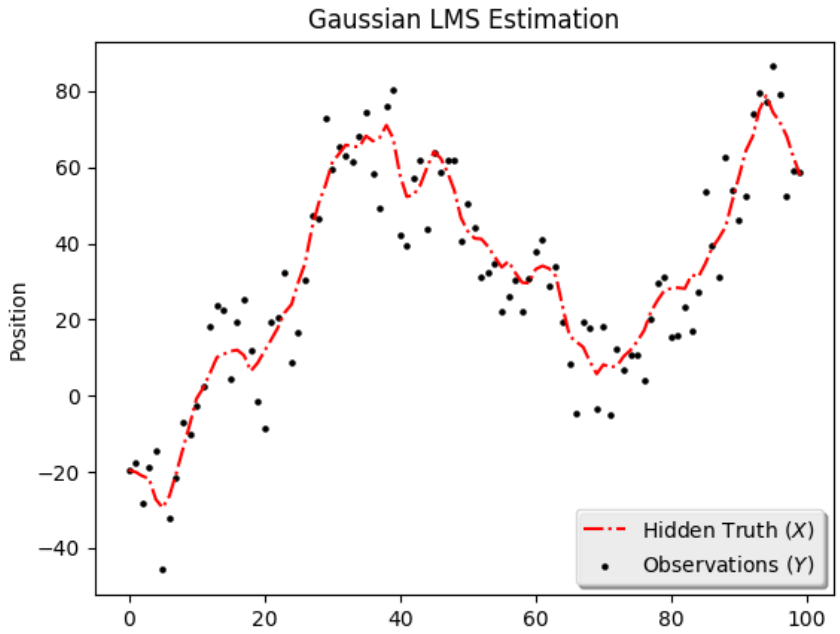
$\alpha > 0$

Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

Tracking a moving vehicle

- Trajectory x_t of vehicle moves with unknown, time-varying velocity v_t
- Discrete-time measurements $Y_t = x_t + W_t$ $\quad W_t \sim N(0, 25)$ independent





- Simplifying/modeling assumption: $x_t = x_{t-1} + v_t$

$$Y_t = x_0 + v_1 + \cdots + v_t + W_t, \qquad t = 1, \ldots, 100$$

$$\theta = (x_0, v_1, \ldots, v_t)$$

Tracking problem: a naive and unsuccessful approach

$Y_t = x_0 + v_1 + \cdots + v_t + W_t, \qquad t = 1, \ldots, 100$

$\theta = (x_0, v_1, \ldots, v_t) \qquad\qquad W_t \sim N(0, 25)$

- Maximize over θ the likelihood $L^\theta(y_1, \ldots, y_{100})$
- For any given θ, the Y_t are independent
  mean $x_0 + \cdots + v_t$; variance 25

$$L^\theta(y_1, \ldots, y_{100}) = c \prod_{i=1}^{100} \exp\left\{ -\frac{1}{2 \cdot 25}(y_t - x_0 - v_1 - \cdots - v_t)^2 \right\}$$

- Solution: can make the exponent equal to zero!

  $\hat{x}_0 + \hat{v}_1 + \cdots + \hat{v}_t = y_t \qquad\qquad \hat{x}_t = y_t$

  – does not extract any useful information

- Must bring in "beliefs" that v_t is not arbitrary $\rightarrow$ go Bayesian!

Introduce "beliefs" on the velocity

- model velocity as a random variable V_t
  - assume $V_t \sim N(0, \sigma^2)$, independent
  - also assume $X_0 \sim N(0, 100)$
  - assume all r.v.s, $X_0, V_1, \ldots, V_{100}, W_1, \ldots, W_{100}$ are independent

$$Y_t = X_0 + V_1 + \cdots + V_t + W_t, \quad t = 1, \ldots, 100$$

$$\Theta = (X_0, V_1, \ldots, V_t) \qquad Y = (Y_1, \ldots, Y_{100})$$

- Conditioned on $X_0 = x_0, V_1 = v_1, \ldots, V_t = v_t$,
  the Y_t are independent, mean $x_0 + \cdots + v_t$; variance 25

$$f_{Y|\Theta}(y \,|\, \theta) = c \prod_{i=1}^{100} \exp\left\{ -\frac{1}{2 \cdot 25}(y_t - x_0 - v_1 - \cdots - v_t)^2 \right\}$$

The Bayes' rule gives a posterior distribution

$$f_{\Theta|Y}(\theta\,|\,y) = \frac{f_\Theta(\theta)\cdot f_{Y|\Theta}(y\,|\,\theta)}{f_Y(y)}$$

$$f_{Y|\Theta}(y\,|\,\theta) = c\prod_{i=1}^{100}\exp\left\{-\frac{1}{2\cdot 25}(y_t - x_0 - v_1 - \cdots - v_t)^2\right\}$$

$$f_\Theta(\theta) = c'\exp\left\{-\frac{x^2}{2\cdot 100}\right\}\cdot\prod_{i=1}^{100}\exp\left\{-\frac{v_t^2}{2\cdot\sigma^2}\right\}$$

- For a given/fixed y, focus on dependence on θ:

  $f_{Y|\Theta}(y\,|\,\theta) = \text{constant}\cdot\exp\left\{-\text{quadratic}(\theta)\right\}$

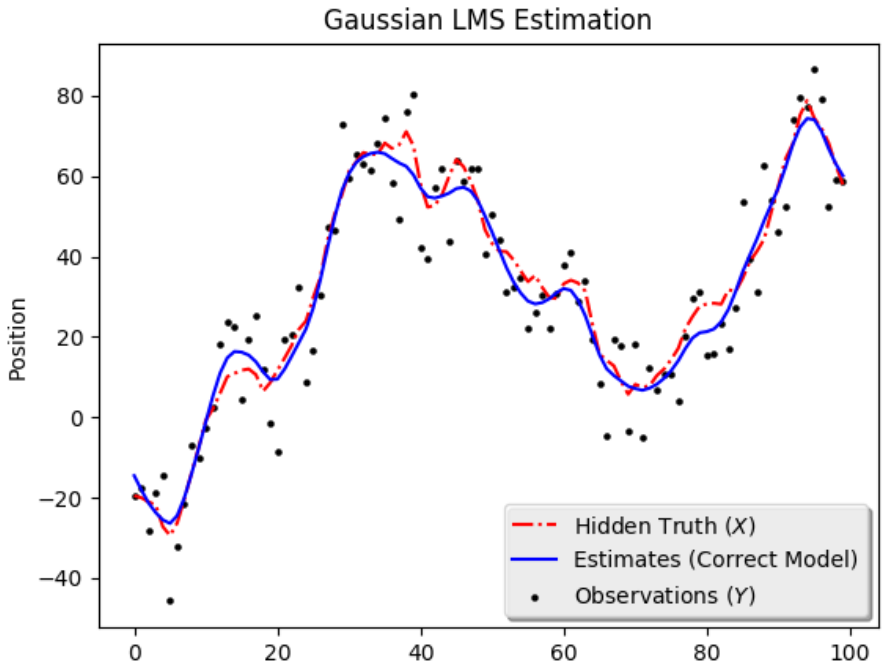
  (such a joint PDF is called a "multivariate normal")

- Optimize the quadratic (linear system for θ) $\rightarrow$ MAP estimate
  - because quadratic functions are symmetric,
    MAP estimate $= \mathbb{E}[\Theta\,|\,Y = y]$

How well does it work?

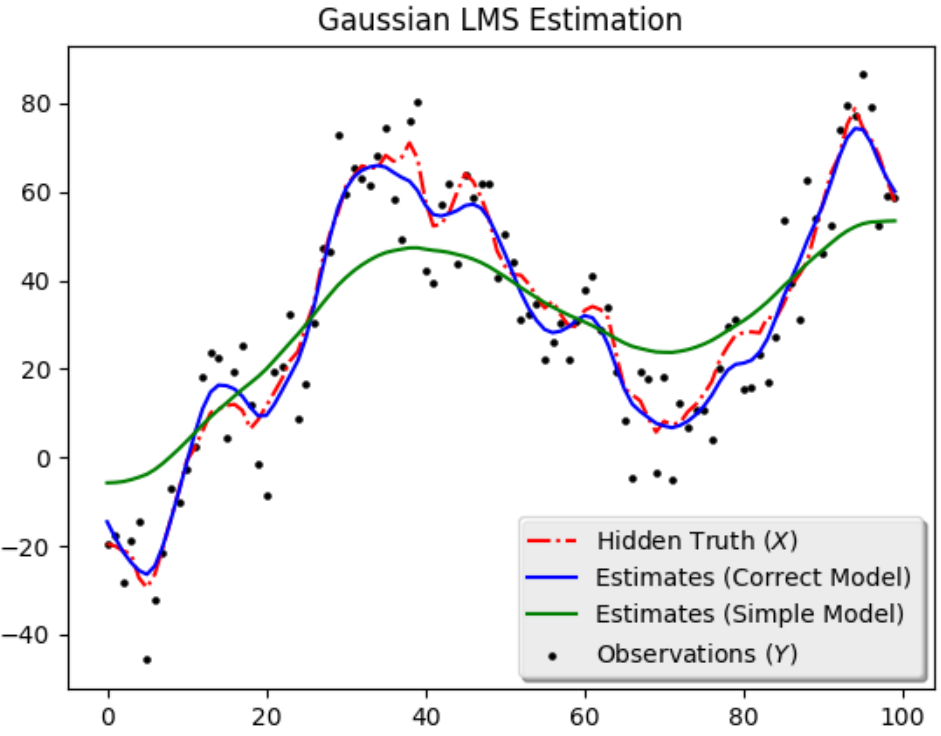
- Ground truth: data were generated from a more complex/realistic model that does not allow sudden velocity changes

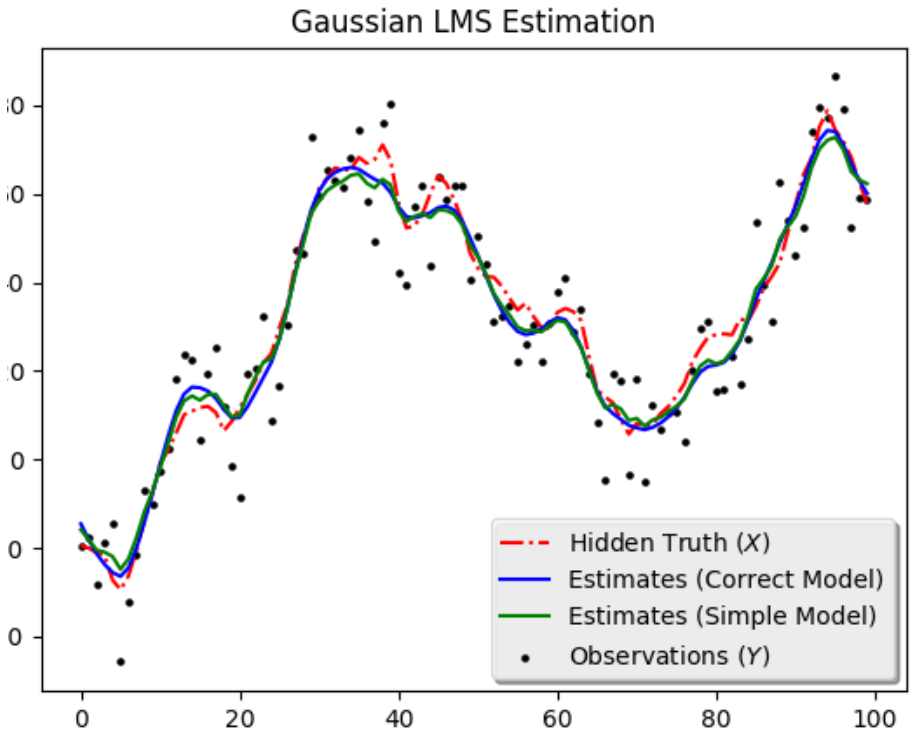
Estimates if we had worked with the true model



How well does it work?

- Estimates based on our (incorrect) model, for different choices of σ^2





- With a good choice of σ, our incorrect model does pretty well
- But how do we set σ?

– prior experience; experimentation with an earlier data set where the ground tuth was known; etc.

Conclusion

- Bayesian inference/estimation can be powerful
  - complete/unique/unambiguous answers to every question

- it is quite tractable (reduces to linear algebra) if:
  - we are dealing with linear normal models, or
  - we restrict to linear estimators

- Difficulties
  - can be computationally intractable
  - it relies on models $\mathbb{P}_{X|\Theta}$, $\mathbb{P}_{\Theta}$
    we may need to use "earlier" data to calibrate these models

- Using "fake," simplified models often gives very good results
  - "all models are wrong, some models are useful"