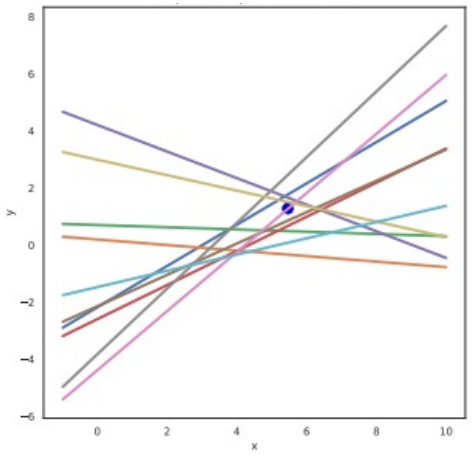
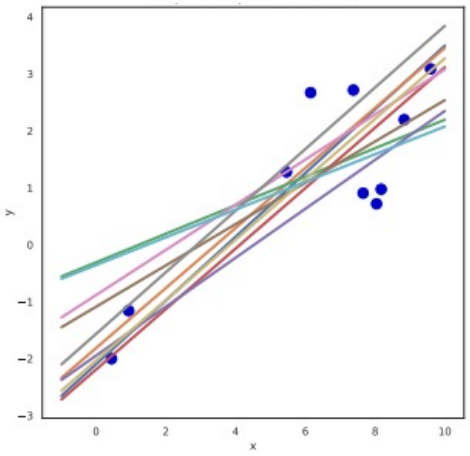
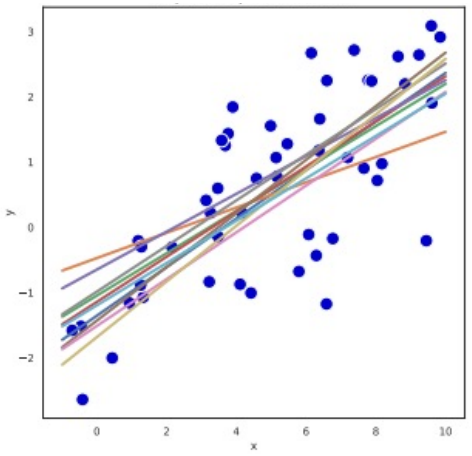
Machine Learning

Bayesian Regression

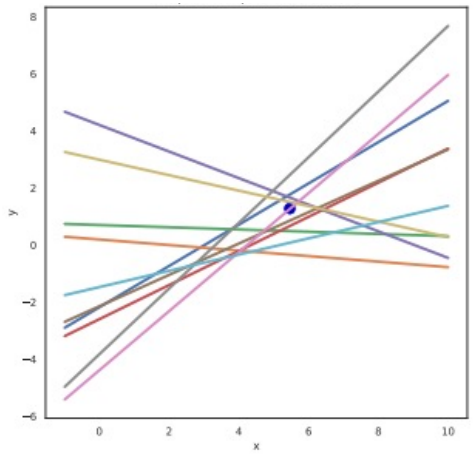
Jian Liu

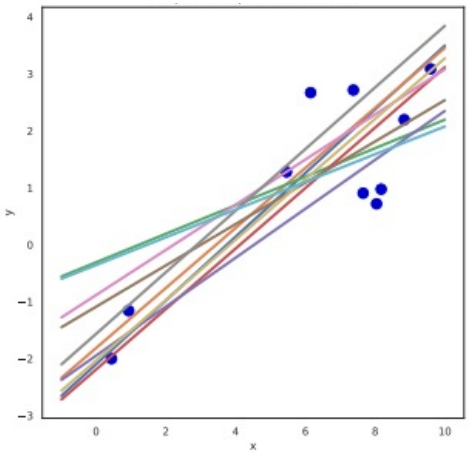


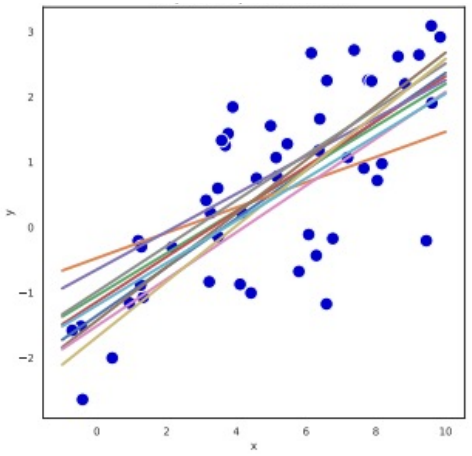




Noisy Targets – Target Distribution







Noisy Targets – Target Distribution

Both convey probabilistic aspects of $\mathbf{x}$ and y

The target distribution $P(y \mid \mathbf{x})$

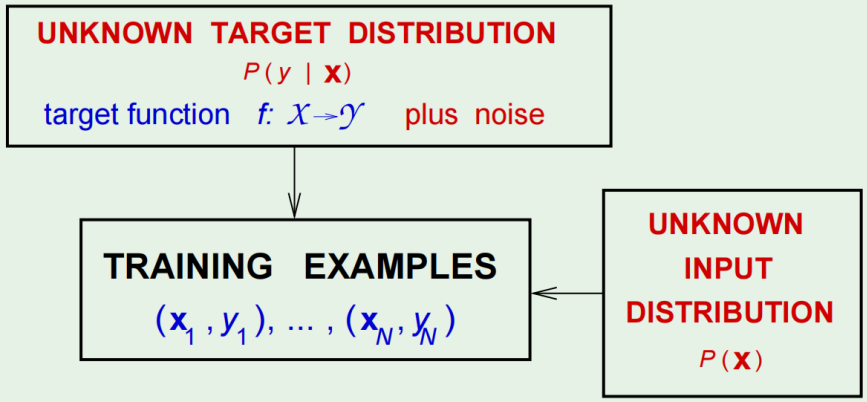
is what we are trying to learn

The input distribution $P(\mathbf{x})$

quantifies relative importance of $\mathbf{x}$

Merging $P(\mathbf{x})P(y|\mathbf{x})$ as $P(\mathbf{x}, y)$

mixes the two concepts



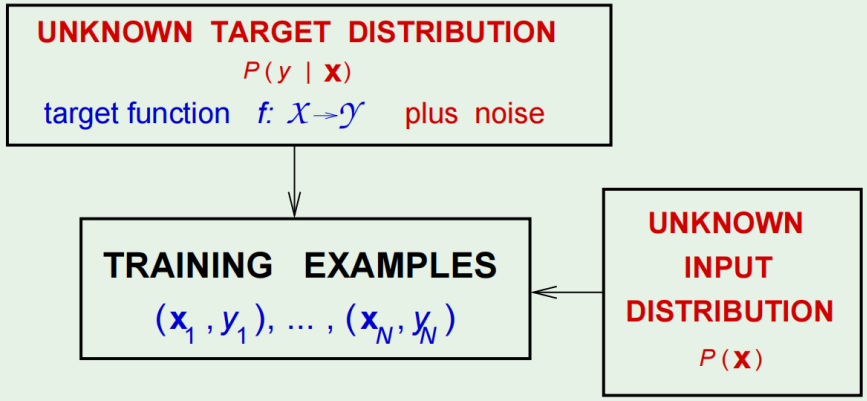
Probabilistic Machine Learning

Both convey probabilistic aspects of $\mathbf{x}$ and y

The target distribution $P(y \mid \mathbf{x})$
is what we are trying to learn

The input distribution $P(\mathbf{x})$
quantifies relative importance of $\mathbf{x}$

Merging $P(\mathbf{x})P(y|\mathbf{x})$ as $P(\mathbf{x}, y)$
mixes the two concepts



Probabilistic Machine Learning

- Key difference in Frequentist vs Bayesian paradigms:
  - Frequentist: model parameters θ are fixed; computed using some *estimator* such as MLE
  - Confidence in estimates for θ evaluated through multiple experiments (cross-validation) for different data sets
  - Bayesian: model parameters θ are random variables
  - Here, there is only one data set (actually observed) and uncertainty in θ is expressed as a probability distribution over θ

Probabilistic Machine Learning
Bayes' Law

- Using the sum and product rules of probability for discrete values of θ (or events):

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}$$

  where, the marginal probability $P(X) = \sum_\theta P(X \mid \theta)P(\theta)$
- For continuous variables, i.e. all possible values of θ we can do the same:

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)}$$

  except now, $p(\theta)$ is a continuous prior distribution rather than discrete probabilities
- Now the marginal probability density function $p(X) = \int p(X|\theta)p(\theta)\partial\theta$
- So we can say: *posterior* $\propto$ *likelihood* $\times$ *prior*

Linear Regression: OLS

$$\hat{y}_i(x_i, \boldsymbol{\omega}) = \sum_{j=0}^{M} \omega_j \phi_j(x_i) = \boldsymbol{\omega}^T \boldsymbol{\phi}(x_i)$$

- By defining design matrix $\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_M(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_M(x_2) \\ \vdots & \vdots & & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_M(x_N) \end{pmatrix}$
- $\boldsymbol{\omega}_{\text{OLS}}$ is given by solving: $\boldsymbol{\Phi}^T \mathbf{y} = \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{\omega}$ The normal equation

$$\boldsymbol{\omega}_{OLS} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{y}$$

Linear Regression: Probabilistic View

- Consider again the linear relationship between target y & predictor x:

$$\mathbf{y} = \hat{y}(\mathbf{x}, \boldsymbol{\omega}) + \epsilon$$

  where, ϵ is some *unexplained* noise; $\hat{y}(\cdot)$ is a linear combination of basis functions
- Lets assume ϵ is a univariate Gaussian variable with zero mean and precision $\beta = \frac{1}{\sigma^2}$
  - β used for convenience of notation
- Then the probability distribution of a target y conditioned on $\mathbf{x}, \boldsymbol{\omega}$ is given by:

$$p(y \mid \mathbf{x}, \boldsymbol{\omega}, \beta) = \mathcal{N}(y \mid \hat{y}(\mathbf{x}, \boldsymbol{\omega}), \beta^{-1})$$

MLE (Maximum Likelihood Estimation)

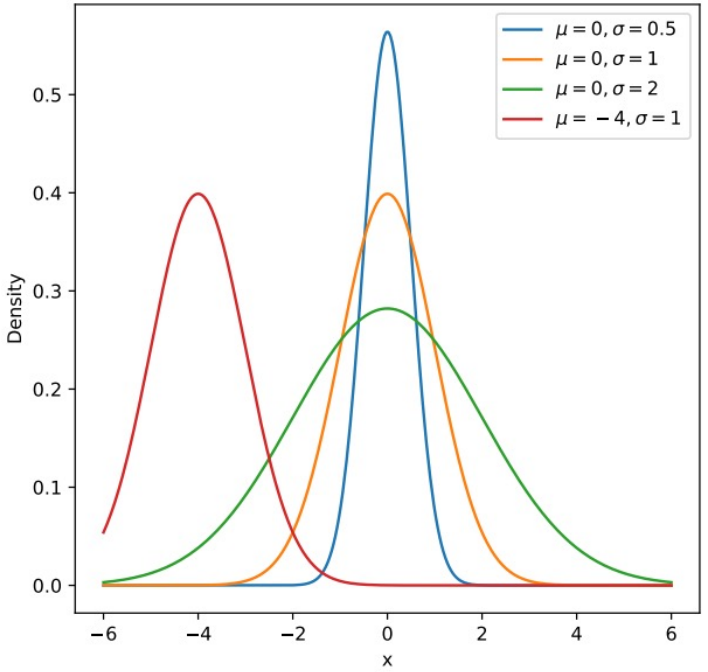
Recap: Univariate Gaussian distribution

- Occurs often in Nature: e.g. height, IQ
- Fixed length sums of variables of any distribution follow approximate Gaussian: Central Limit Theorem
- μ: position of centre; σ: width of distribution

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{1}$$

$$\mathbb{E}[\mathcal{N}(x \mid \mu, \sigma^2)] = \mu \tag{2}$$

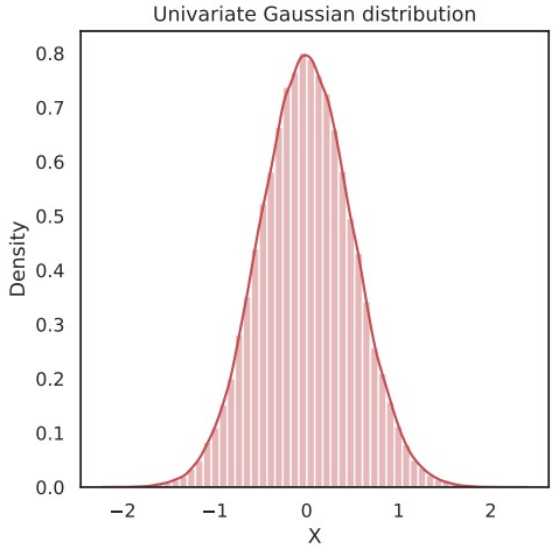
$$\text{var}[\mathcal{N}(x \mid \mu, \sigma^2)] = \sigma^2$$



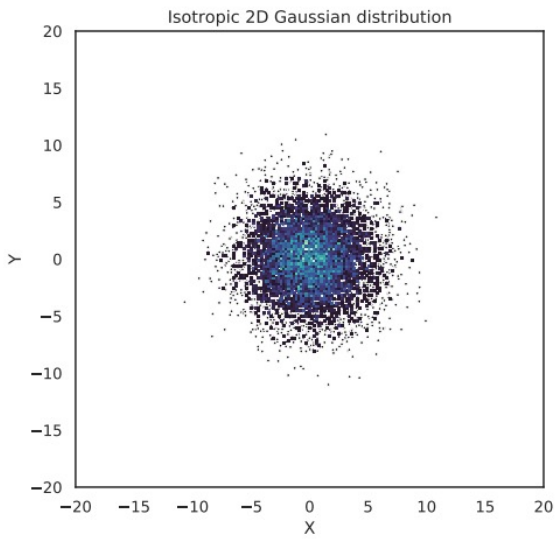
MLE (Maximum Likelihood Estimation)

Recap: Multivariate Gaussian Distribution

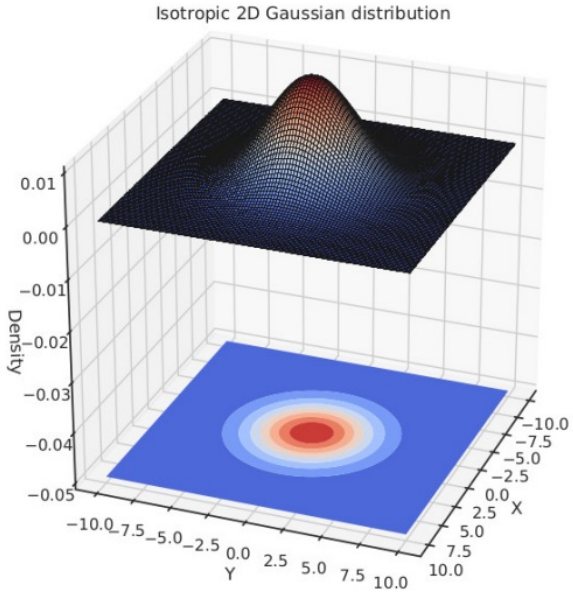
- Ubiquity of the Gaussian distribution can be explained by the Central Limit Theorem
- This holds for Univariate and Multivariate Gaussian distributions
- What is a multivariate distribution?



(a) Univariate Gaussian distribution



(b) Samples from bivariate Gaussian



(c) Density of bivariate Gaussian

MLE (Maximum Likelihood Estimation)

Recap: Multivariate Gaussian Distribution

- For an N-dim vector $\mathbf{x}$ the multivariate Gaussian distribution is given by:

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{N/2}} \frac{1}{|\det(\boldsymbol{\Sigma})|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

- Here, $\boldsymbol{\mu}$ is an N-dim mean vector and $\boldsymbol{\Sigma}$ is an $N \times N$ covariance matrix
- For example for a bivariate (2D) Gaussian: $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{xx}^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy}^2 \end{bmatrix}$

MLE (Maximum Likelihood Estimation)

- Parameters to estimate when fitting a multivariate Gaussian to data are: $\boldsymbol{\mu}, \boldsymbol{\Sigma}$
- Lets look at deriving MLE for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$
- Given a dataset $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ of N *i.i.d* data points each forming a row of the matrix,
- Using the product rule we define the likelihood function as:

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} p(\mathbf{x_i}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- And we know that for a single data point $\mathbf{x}_i$ we have:

$$p(\mathbf{x_i}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{N/2}} \frac{1}{\mid \det(\boldsymbol{\Sigma}) \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})\right\}$$

MLE (Maximum Likelihood Estimation)

- To derive MLE for $\boldsymbol{\mu}$ & $\boldsymbol{\Sigma}$ we need to maximise the likelihood function w.r.t each parameter
- I.e. we need to solve:

$$\frac{\partial}{\partial \boldsymbol{\mu}} p(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid_{\mu=\mu_{ML}} = 0$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} p(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid_{\Sigma=\Sigma_{ML}} = 0$$

- Easier to work with the log-likelihood function ($\boldsymbol{\mathcal{L}}$):

$$\boldsymbol{\mathcal{L}} = \ln p(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln \det(\boldsymbol{\Sigma}) - \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

MLE (Maximum Likelihood Estimation)

- Considering just the terms in $(\mathcal{L})$ dependent on $\boldsymbol{\mu}$ & $\boldsymbol{\Sigma}$:

$$\mathcal{L}(\boldsymbol{\mu}) = \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$\mathcal{L}(\boldsymbol{\Sigma}) = -\frac{N}{2} \ln \det(\boldsymbol{\Sigma}) - \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

- Solving $\frac{\partial \mathcal{L}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = 0$ & $\frac{\partial \mathcal{L}(\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = 0$ gives:

$$\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

Linear Regression: MLE & OLS

- Given a data set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1\ldots N}$ & targets $\mathbf{y} = \{y_i\}_{i=1\ldots N}$, assuming data are *i.i.d*,
- We can formulate the likelihood function as a function of $\boldsymbol{\omega}, \beta$:

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\omega}, \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \boldsymbol{\omega}^T \boldsymbol{\phi}(x_i), \beta^{-1})$$

where we have used the previous result: $\hat{y}(x_i, \boldsymbol{\omega}) = \boldsymbol{\omega}^T \boldsymbol{\phi}(x_i)$

- But we have seen this form before for MLE of multivariate Gaussians...

we can express the log-likelihood $\mathcal{L}$ as:

$$\boldsymbol{\mathcal{L}} = \ln p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\omega}, \beta) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln \beta - \frac{1}{2} \sum_{i=1}^{N} \beta (y_i - \boldsymbol{\omega}^T \boldsymbol{\phi}(x_i))^2$$

Linear Regression: MLE & OLS

- Looking at the quadratic term in $\mathcal{L}$:

$$\mathcal{L} = \ln p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\omega}, \beta) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln\beta - \frac{\beta}{2}R(\boldsymbol{\omega})$$

$$R(\boldsymbol{\omega}) = \sum_{i=1}^{N}(y_i - \boldsymbol{\omega}^T\boldsymbol{\phi}(x_i))^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i(x_i, \boldsymbol{\omega}))^2$$

- But this is the same as the sum of squared residuals in OLS!
- $\boldsymbol{\omega}$ which maximises $\mathcal{L}$ also minimises R: $\arg\max_{\boldsymbol{\omega}} \mathcal{L} \equiv \arg\min_{\boldsymbol{\omega}}(-\mathcal{L})$

$$\arg\min_{\boldsymbol{\omega}}(-\mathcal{L}(\boldsymbol{\omega})) = \arg\min_{\boldsymbol{\omega}} \sum_{i=1}^{N}(y_i - \hat{y}_i(x_i, \boldsymbol{\omega}))^2 = \arg\min_{\boldsymbol{\omega}} R(\boldsymbol{\omega})$$

- So $\boldsymbol{\omega}_{MLE} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{y}$ under additive Gaussian noise assumption

$$\boldsymbol{\omega}_{OLS} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{y}$$

MLE = OLS

Linear Regression: MLE & OLS

Regression

Optimization:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$$

Solution:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Assumption: None

Maximum-Likelihood

Optimization:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \left(\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2\right\}$$

Solution:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

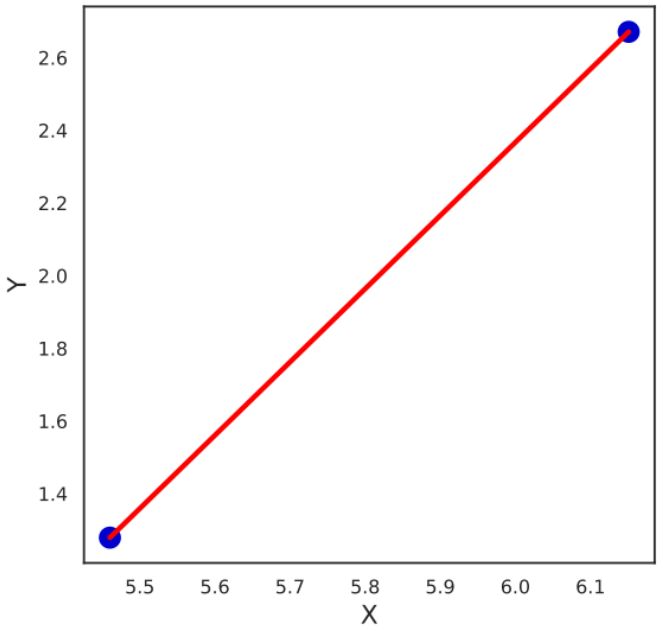
Assumption: $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta} \sim \text{Gaussian}(0, \sigma^2\boldsymbol{I})$

Linear Regression: Bayesian perspective

- Consider linear regression with a polynomial of degree $= 1$; $\hat{y}_i = \omega_0 + \omega_1 x_i$
- If we have just two predictors and their targets: $\{x_i, y_i\}_{i=1,2}$, we can fit a line exactly

  - Here we have two unknowns and two observations
  - Can determine the slope (ω_1) and intercept (ω_0) just using the observations
  - We can solve for ω_1 as follows and then estimate ω_0:

$$\omega_1 = \frac{y_2 - y_1}{x_2 - x_1} \tag{13}$$

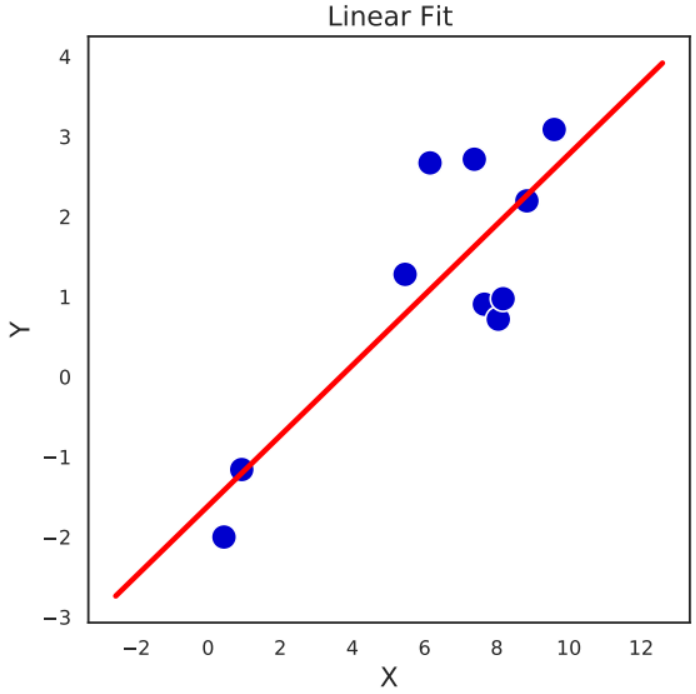


Linear Regression: Bayesian perspective

- But what if we had more than two points?
  - Then we have an *overdetermined* system
  - We can use OLS or assume a noise model
  - Then system of linear equations are:

$$y_i = \omega_0 + \omega_1 x_i + \epsilon_i \tag{14}$$

    where, we can assume $\epsilon \sim \mathcal{N}(0, \sigma^2)$
  - And this leads to $\boldsymbol{\omega}_{MLE} \equiv \boldsymbol{\omega}_{OLS}$



Linear Regression: Bayesian perspective

- What if we have an *underdetermined* system? E.g. just one observation
- We have two unknowns and one observation: $y_1 = \omega_0 + \omega_1 x_1 + \epsilon_1$
- What can we do if we want to fit a line to this?

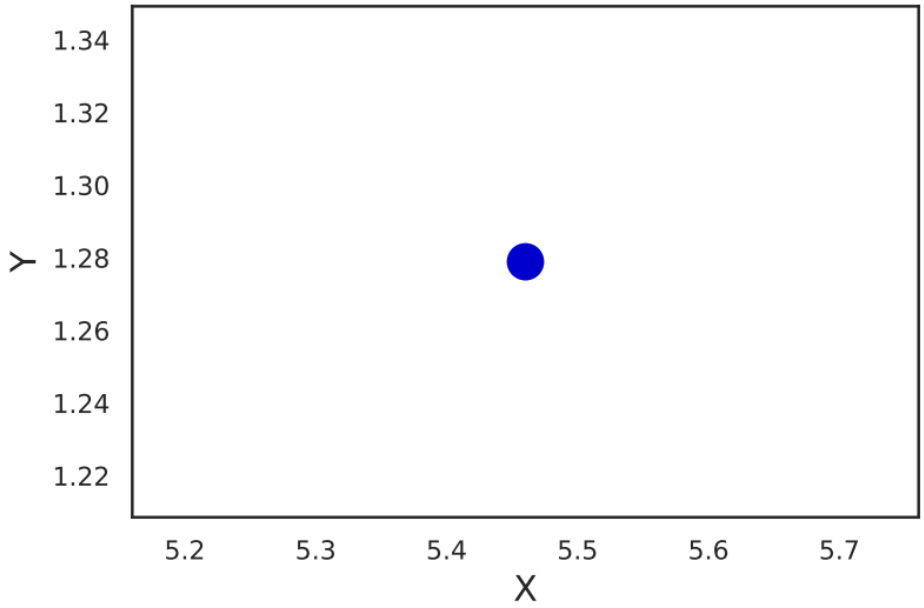


Figure: Just one observations (x_1, y_1)

Linear Regression: Bayesian perspective

- What if we have an *underdetermined* system? E.g. just one observation
- We have two unknowns and one observation: $y_1 = \omega_0 + \omega_1 x_1 + \epsilon_1$
- We have an infinite number of possible solutions..
- How can we represent a family of ω_0?

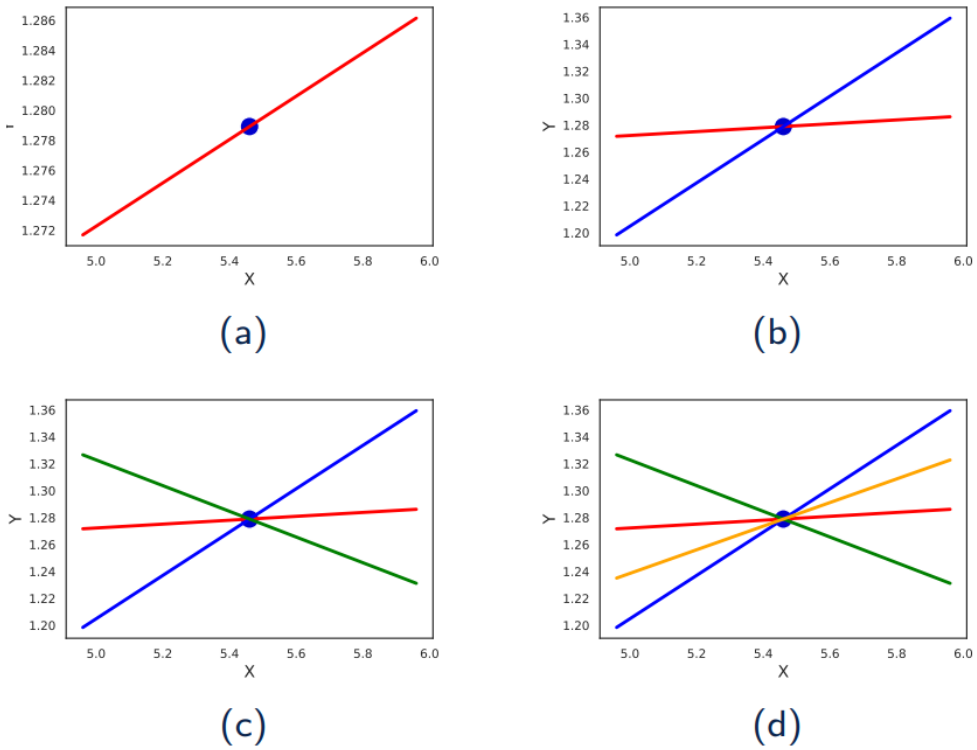


Figure: (a): $\omega_0 = 1.2$, (b): $\omega_0 = 0.4$, (c): $\omega_0 = 1.8$, (d): $\omega_0 = 0.8$

Linear Regression: Bayesian perspective

- We can make an assumption for the distribution of the unknown parameter ω_0!
- And that is Bayesian Inference!
- Given a data set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1\ldots N}$ & targets $\mathbf{y} = \{y_i\}_{i=1\ldots N}$, assuming data are *i.i.d*,
- As previously the likelihood function is given by:

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\omega}, \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \boldsymbol{\omega}^T \boldsymbol{\phi}(x_i), \sigma^2)$$

- We tried to *maximise the likelihood* with respect to $\boldsymbol{\omega}$
- Bayesian perspective: based on some *a priori* belief about $\boldsymbol{\omega}$, integrate across $\boldsymbol{\omega}$
- I.e. instead of fixed $\boldsymbol{\omega}$ we look at expectation of likelihood for a range of plausible $\boldsymbol{\omega}$

Linear Regression: Bayesian perspective

- For Bayesian inference we first define a prior distribution on the unknowns/parameters ($\boldsymbol{\omega}$)
- Prior represents our belief about $\boldsymbol{\omega}$ before we see the data
- Linear regression: let's consider a Gaussian prior on the intercept (ω_0) (ignore ω_1 for now):

$$p(\omega_0) \sim \mathcal{N}(0, \alpha)$$

- Given N i.i.d observations $\{x_i, y_i\}_{i=1\ldots N}$ and $p(\omega_0)$ the posterior distribution of ω_0 is:

$$p(\omega_0 \mid \mathbf{y}, \mathbf{x}, \omega_1, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \omega_1 x_i - \omega_0)^2 - \frac{\omega_0^2}{2\alpha}\right\}$$

$$\ln p(\omega_0 \mid \mathbf{y}, \mathbf{x}, \omega_1, \sigma^2) = \left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \omega_1 x_i - \omega_0)^2 - \frac{\omega_0^2}{2\alpha}\right\} + const$$

Linear Regression: Bayesian perspective

- Once again by completing the square we find:

$$\ln p(\omega_0 \mid \mathbf{y}, \mathbf{x}, \omega_1, \sigma^2) = -\frac{1}{\sigma^2_{post}}(\omega_0 - \mu_{post})^2 + const$$

$$p(\omega_0 \mid \mathbf{y}, \mathbf{x}, \omega_1, \sigma^2) \sim \mathcal{N}(\mu_{post}, \sigma^2_{post})$$

$$\mu_{post} = \frac{N\alpha}{N\alpha + \sigma^2}\mu_{ML}, \qquad \sigma^2_{post} = \frac{\sigma^2\alpha}{N\alpha + \sigma^2}$$

- Where μ_{ML} is the MLE for the intercept ω_0 (refer to C. Bishop's book pg. 142)
- For MLE we estimated all regression weights jointly
- We can infer the posterior distribution over all weights by assuming a prior $p(\omega)$ over all

Linear Regression: Bayesian perspective

- Previously, we assumed a prior over the intercept for a linear regression function
- What about a prior over all regression weights?
- Given a data set N i.i.d observations: $\mathbf{y} = \hat{y}(\mathbf{x}, \boldsymbol{\omega}) + \boldsymbol{\epsilon}$

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\omega}, \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \boldsymbol{\omega}^T \boldsymbol{\phi}(x_i), \beta^{-1})$$

$$p(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega} \mid \mathbf{0}, \alpha^{-1}\mathbb{I})$$

- Then posterior distribution over weights is given by:

$$p(\boldsymbol{\omega} \mid \mathbf{y}, \mathbf{x}, \beta) \propto p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\omega}, \beta) p(\boldsymbol{\omega})$$

- We will treat the precision β as a known constant - but this may be inferred as well

Linear Regression: Bayesian perspective

- We know if $p(\boldsymbol{\omega})$ is a conjugate prior then $p(\boldsymbol{\omega} \mid \mathbf{y}, \mathbf{x})$ is a Gaussian distribution
- As seen previously, given $p(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega} \mid \mathbf{0}, \alpha^{-1}\mathbb{I})$:

$$\ln p(\boldsymbol{w} \mid \mathbf{y}, \mathbf{x}) = -\frac{\beta}{2}\sum_{i=1}^{N}\left\{y_i - \boldsymbol{w}^T\phi(x_i)\right\}^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{const}$$

- Once again by *completing the square*:

$$p(\boldsymbol{\omega} \mid \mathbf{y}, \mathbf{x}) = \mathcal{N}(\boldsymbol{\omega} \mid \mathbf{m}_{post}, \mathbf{S}_{post})$$

$$\mathbf{m}_{post} = \beta\mathbf{S}_{post}\boldsymbol{\Phi}^T\mathbf{y}, \qquad \mathbf{S}_{post}^{-1} = \alpha\mathbb{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

where, $\boldsymbol{\Phi}$ is the design matrix, $\mathbf{m}_{post}, \mathbf{S}_{post}$ are the mean and covariance of the posterior distribution for $\boldsymbol{\omega}$

Linear Regression: Bayesian perspective

- MLE for linear regression: $\boldsymbol{\omega}_{MLE} = \boldsymbol{\Phi}^{\dagger}$
- MLE gives point estimates for $\boldsymbol{\omega}$, but a predictive distribution in *data space*
- Remember:
  $p(y_i \mid x_i, \boldsymbol{\omega}, \beta) = \mathcal{N}(y_i \mid \boldsymbol{\omega}^T \phi(x_i), \beta^{-1})$
- Ex: Fitting degree=1 polynomial; $\beta = 1.0$
- Predictive distribution is:
  $\mathcal{N}(y \mid (\omega_0 + \omega_1 x), 1.0)$

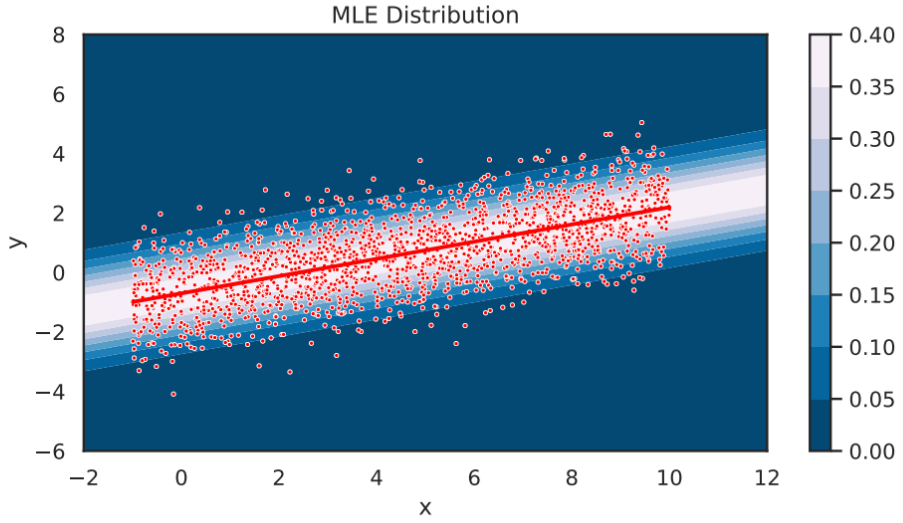


Figure: Polynomial degree=1 fit to data; contour plot depicts predictive distribution

- What do you notice about the MLE distribution?

Linear Regression: Bayesian perspective

- Bayesian inference for linear regression: posterior distribution of weights $p(\boldsymbol{\omega} \mid \mathbf{y}, \mathbf{x})$
- I.e. we have a predictive distribution for each possible $\boldsymbol{\omega}$ from $p(\boldsymbol{\omega} \mid \mathbf{y}, \mathbf{x})$
- Consider $N = 10$ i.i.d observations, $\beta = 1.0$, prior $p(\boldsymbol{\omega}) \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbb{I})$
- We infer $p(\boldsymbol{\omega} \mid \mathbf{y}, \mathbf{x}) = \mathcal{N}(\boldsymbol{\omega} \mid \mathbf{m}_{post}, \mathbf{S}_{post})$: What do you notice?

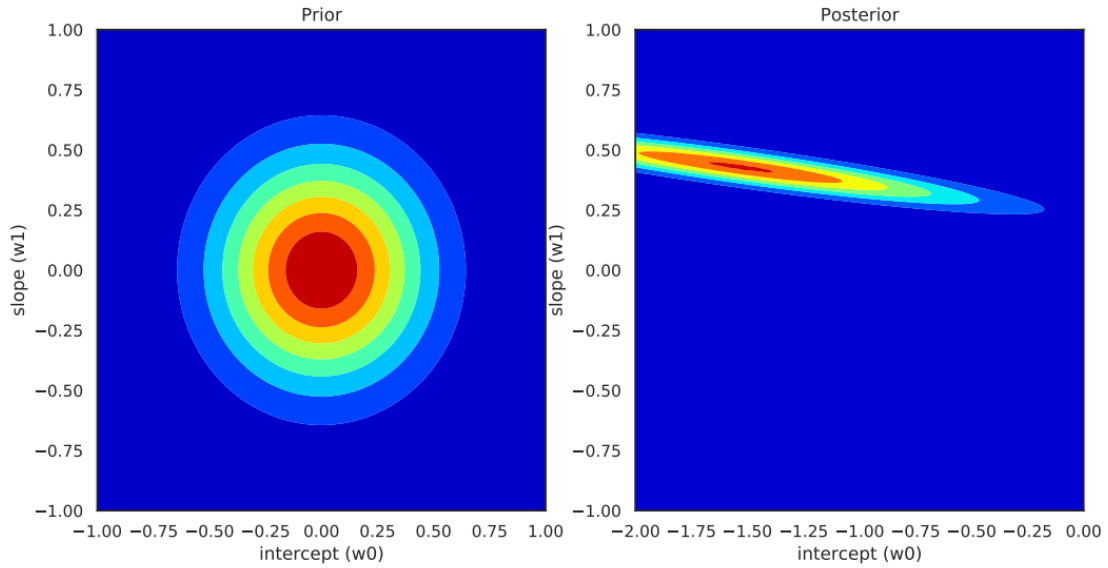
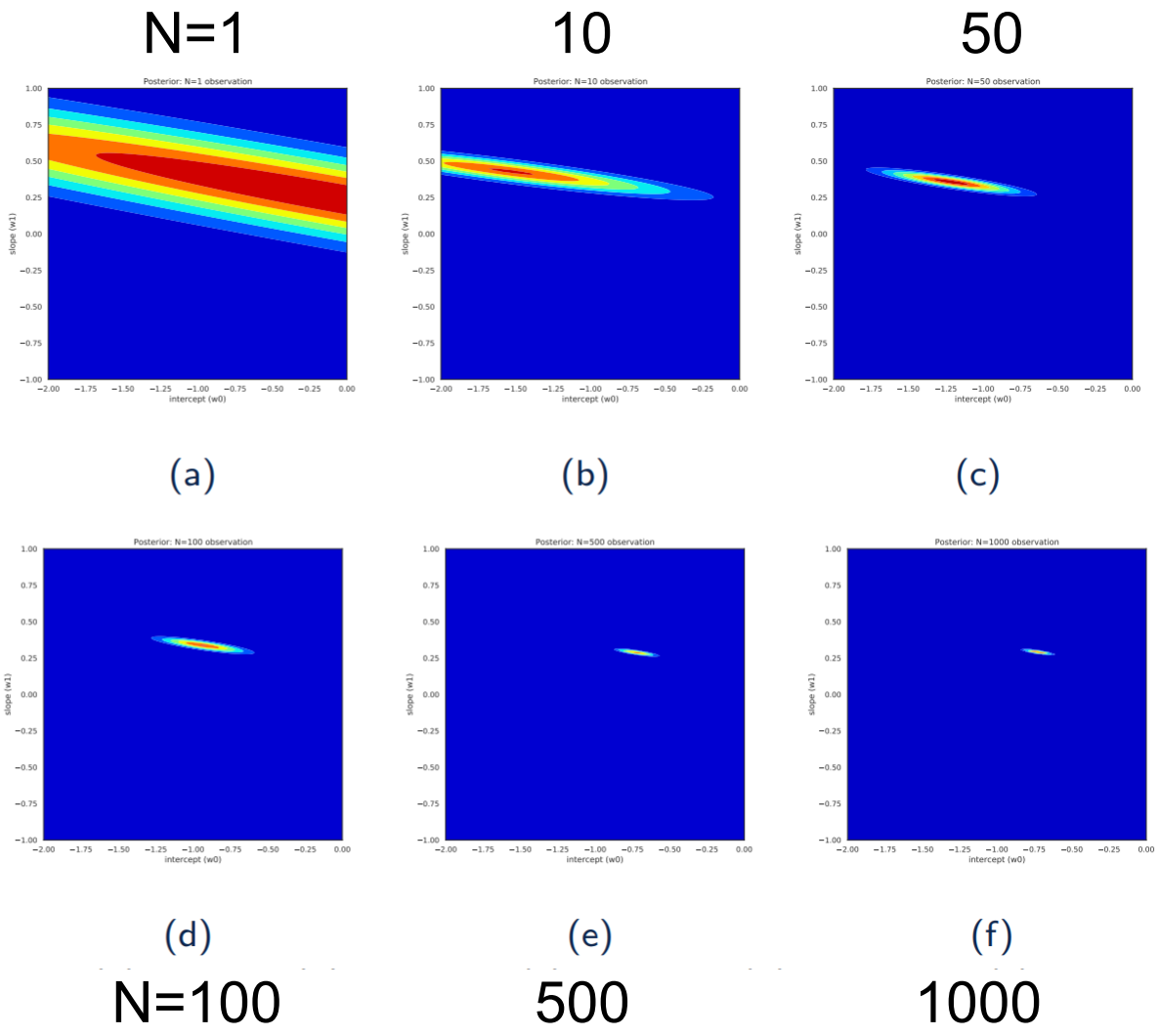


Figure: Left: Prior distribution, $\alpha = 10.0$; Right: Posterior distribution after $N = 10$ observations

Linear Regression: Bayesian perspective

- Let's see what the impact of varying the number of observations is on the posterior
- What do you notice?
- x-axis: intercept (ω_0); y-axis: slope (ω_1)



Linear Regression: Bayesian perspective

- *Model uncertainty* reduces with increase in observations
- What about the predictive distribution?
- Let's sample 10 instances of ω from $\mathcal{N}(\mathbf{m}_{post}, \mathbf{S}_{post})$
- What do you notice?

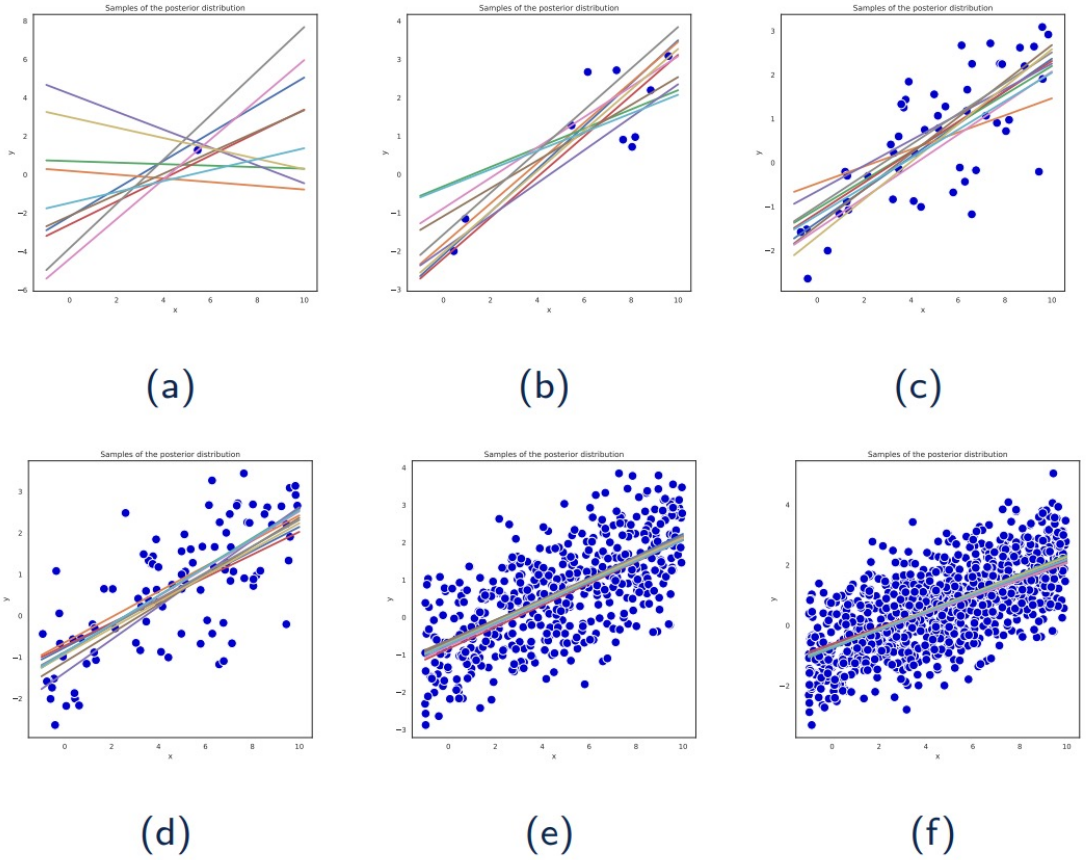


Figure: (a) $N = 1$; (b) $N = 10$; (c) $N = 50$; (d) $N = 100$; (e) $N = 500$; (f) $N = 1000$

Summary

- OLS = MLE (Point estimation)
- Bayesian linear regression models (Distribution of model parameters)
- Data are too few or too noisy