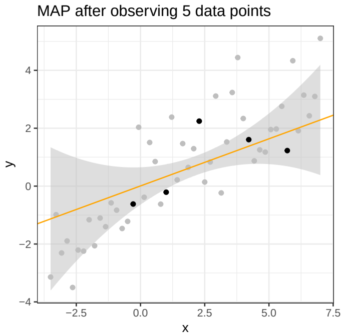
Introduction to Machine Learning

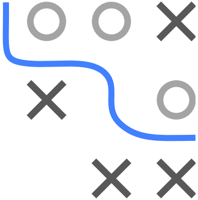
Gaussian Processes
Bayesian Linear Model



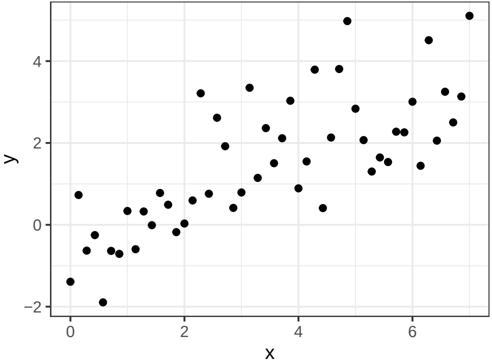
Learning goals

- Know the Bayesian linear model
- The Bayesian LM returns a (posterior) distribution instead of a point estimate
- Know how to derive the posterior distribution for a Bayesian LM

DATA SITUATION



- $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$: i.i.d. training set from some unknown distribution



- $\mathbf{X} \in \mathbb{R}^{n \times p}$: design matrix, where i-th row contains vector $\mathbf{x}^{(i)}$
- $\mathbf{y} = \left(y^{(1)}, \ldots, y^{(n)}\right)^\top$

BAYESIAN LINEAR MODEL REVISITED

- Standard linear regression model for i-th observation, with $\boldsymbol{\theta} \in \mathbb{R}^p$ fixed but unknown

$$y^{(i)} = f\left(\mathbf{x}^{(i)}\right) + \epsilon^{(i)} = \boldsymbol{\theta}^T \mathbf{x}^{(i)} + \epsilon^{(i)} \quad \forall i$$

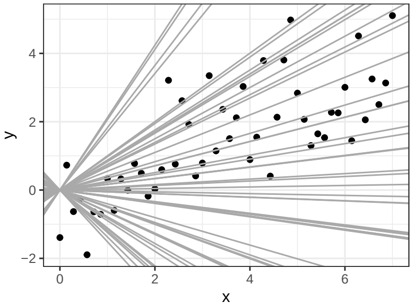
- Assumption: function outputs $f\left(\mathbf{x}^{(i)}\right)$ differ from observed values $y^{(i)}$ by additive, i.i.d. Gaussian noise (ind of $\mathbf{x}$, $\boldsymbol{\theta}$)

$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2) \quad \forall i$$

- Bayesian perspective: $\boldsymbol{\theta}$ also RV with associated (prior) distribution, e.g., $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_p)$

GP PERSPECTIVE

- Weight-space view: prior over $\boldsymbol{\theta}$, function-space view: prior over linear functions
- Example: random lines with intercept 0, slope $\boldsymbol{\theta} \sim \mathcal{N}(0, 1)$



- Random lines = draws from GP with linear kernel
- Collection of RVs $\{f(\mathbf{x}) = \boldsymbol{\theta}\mathbf{x} : \mathbf{x} \in \mathbb{R}\}$
- $f(\mathbf{X})$ is mv Gaussian for any finite input with design matrix $\mathbf{X}$

FROM PRIOR TO POSTERIOR

- Bayes' rule: update prior to posterior belief after observing data

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \frac{\text{likelihood} \cdot \text{prior}}{\text{marginal likelihood}} = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \cdot q(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{X})}$$

- Gaussian family is "self-conjugate": Gaussian prior & Gaussian likelihood $\Rightarrow$ Gaussian posterior

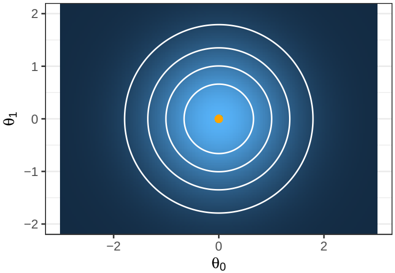
$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{y}, \mathbf{K}^{-1})$$

  with $\mathbf{K} := \sigma^{-2}\mathbf{X}^T\mathbf{X} + \tau^{-2}\boldsymbol{I}_p$

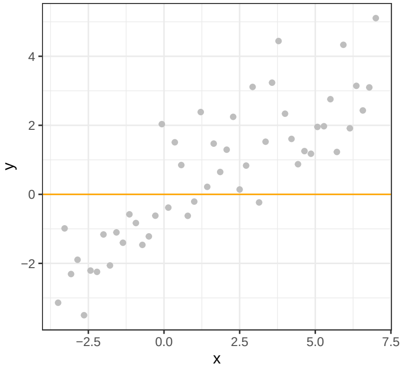
- Intuitively: quantifies posterior (i.e., after seeing data) probability of $\boldsymbol{\theta}$ having generated the observed data

POSTERIOR CONTRACTION

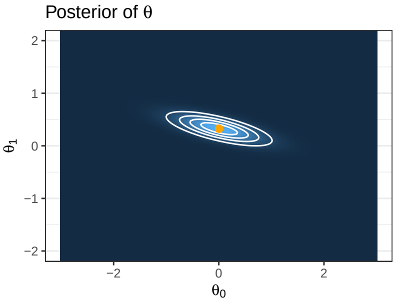
Prior $\theta \sim N(0, 1)$

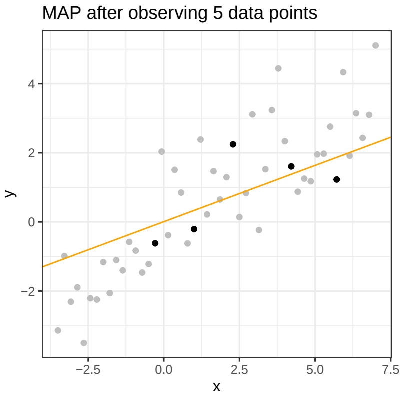


No data points observed

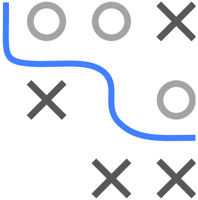


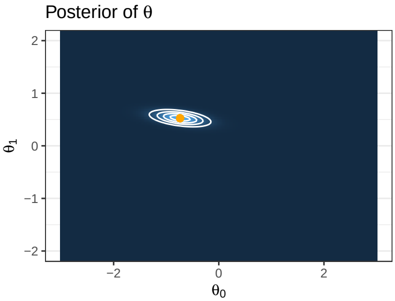
POSTERIOR CONTRACTION

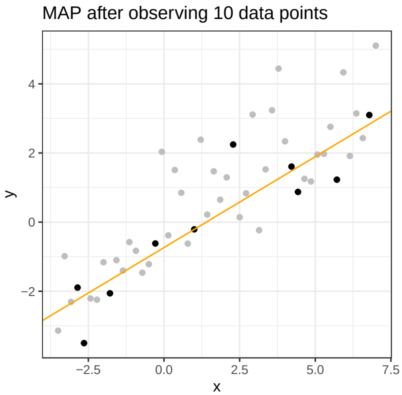




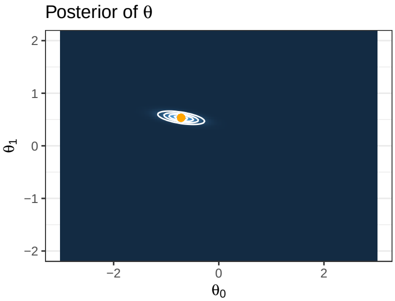
POSTERIOR CONTRACTION

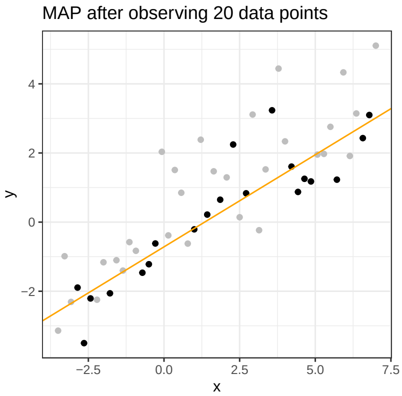






POSTERIOR CONTRACTION





PROOF: GAUSSIANITY OF POSTERIOR

- We want to show that for
  - Gaussian prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_p)$ and
  - Gaussian likelihood $\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}^T\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_n)$

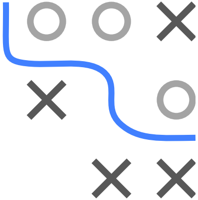
  the resulting posterior is Gaussian: $\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{y}, \mathbf{K}^{-1})$

- Plug in Bayes' rule and keep only terms depending on $\boldsymbol{\theta}$

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) \quad &\propto \quad p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})q(\boldsymbol{\theta}) \quad \propto \quad \exp[-\tfrac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) - \tfrac{1}{2\tau^2}\boldsymbol{\theta}^T\boldsymbol{\theta}] \\
&= \quad \exp\left[-\tfrac{1}{2}(\sigma^{-2}\mathbf{y}^T\mathbf{y} - 2\sigma^{-2}\mathbf{y}^T\mathbf{X}\boldsymbol{\theta} + \sigma^{-2}\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} + \tau^{-2}\boldsymbol{\theta}^T\boldsymbol{\theta})\right] \\
&\propto \quad \exp\left[-\tfrac{1}{2}(\sigma^{-2}\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} + \tau^{-2}\boldsymbol{\theta}^T\boldsymbol{\theta} - 2\sigma^{-2}\mathbf{y}^T\mathbf{X}\boldsymbol{\theta})\right] \\
&= \quad \exp\left[-\tfrac{1}{2}\boldsymbol{\theta}^T\underbrace{(\sigma^{-2}\mathbf{X}^T\mathbf{X} + \tau^{-2}\boldsymbol{I}_p)}_{:=\mathbf{K}}\boldsymbol{\theta} + \sigma^{-2}\mathbf{y}^T\mathbf{X}\boldsymbol{\theta}\right]
\end{aligned}
$$

- Note how this resembles a normal density, except for term in orange
- (No need to worry about normalizing constant $\Rightarrow$ sole purpose: ensure density integrates to total prob of 1)

PROOF: GAUSSIANITY OF POSTERIOR



- Trick: introduce constant c, compensating for added quantities ("creative 0"), s.t. additions will conveniently cancel out with nuisance term

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) &\propto \exp[-\tfrac{1}{2}(\boldsymbol{\theta}-c)^T\mathbf{K}(\boldsymbol{\theta}-c)-c^T\mathbf{K}\boldsymbol{\theta}+\underbrace{\tfrac{1}{2}c^T\mathbf{K}c}_{\text{doesn't depend on } \boldsymbol{\theta}}+\sigma^{-2}\mathbf{y}^T\mathbf{X}\boldsymbol{\theta}] \\
&\propto \exp[-\tfrac{1}{2}(\boldsymbol{\theta}-c)^T\mathbf{K}(\boldsymbol{\theta}-c)-c^T\mathbf{K}\boldsymbol{\theta}+\sigma^{-2}\mathbf{y}^T\mathbf{X}\boldsymbol{\theta}]
\end{aligned}
$$

- Choosing c s.t. $-c^T\mathbf{K}\boldsymbol{\theta}+\sigma^{-2}\mathbf{y}^T\mathbf{X}\boldsymbol{\theta}=0$ leads to $\boldsymbol{\theta} \mid \mathbf{X},\mathbf{y} \sim \mathcal{N}(c, \mathbf{K}^{-1})$
- Using that $\mathbf{K}$ is symmetric, this implies

$$
\begin{aligned}
& \sigma^{-2}\mathbf{y}^T\mathbf{X} = c^T\mathbf{K} \\
\Leftrightarrow \quad & \sigma^{-2}\mathbf{y}^T\mathbf{X}\mathbf{K}^{-1} = c^T \\
\Leftrightarrow \quad & c = \sigma^{-2}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}
$$

- Finally: $\boldsymbol{\theta} \mid \mathbf{X},\mathbf{y} \sim \mathcal{N}(\sigma^{-2}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{y}, \mathbf{K}^{-1}) \quad \square$

POSTERIOR PREDICTIVE DISTRIBUTION

- How does prediction change w.r.t. classical (non-Bayesian) LM?
- Gaussian posterior

$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{y}, \mathbf{K}^{-1})$$

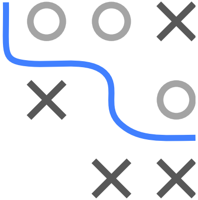
  induces (Gaussian) predictive distribution

- For a new observation $\mathbf{x}_*$ we get

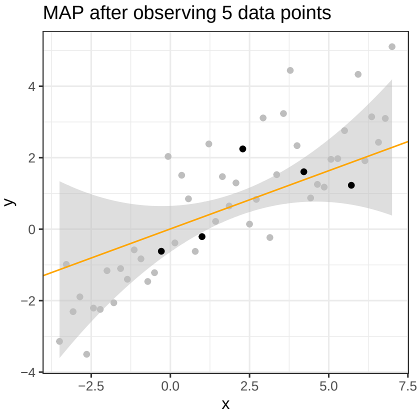
$$y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}(\sigma^{-2}\mathbf{y}^T\mathbf{X}\mathbf{K}^{-1}\mathbf{x}_*, \mathbf{x}_*^T\mathbf{K}^{-1}\mathbf{x}_*)$$

- Intuitively: expectation over all $\boldsymbol{\theta}$-parameterized LMs, weighted according to posterior prob $\leftrightarrow$ classical LM: only max-prob $\boldsymbol{\theta}^{\text{MAP}}$
- Entire distribution with built-in uncertainty quantification!

POSTERIOR MEAN AND VARIANCE

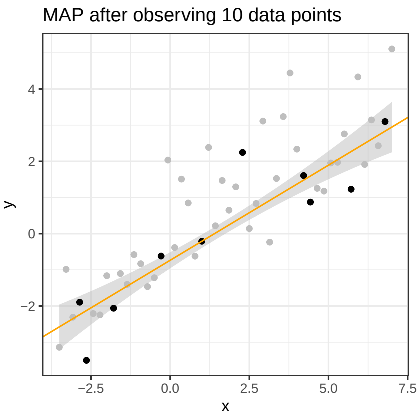


- For every test input $\mathbf{x}_*$, we get a posterior mean (orange) & variance (grey region; $\pm 2\times$ standard deviation)



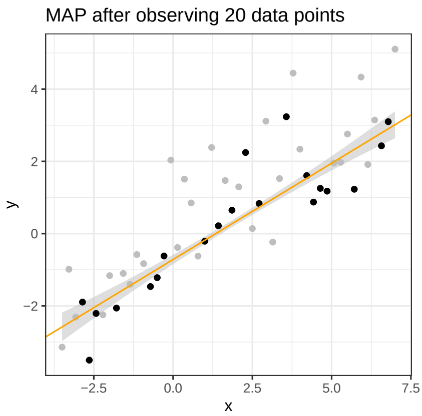
POSTERIOR MEAN AND VARIANCE

- For every test input $\mathbf{x}_*$, we get a posterior mean (orange) & variance (grey region; $\pm 2\times$ standard deviation)



POSTERIOR MEAN AND VARIANCE

- For every test input $\mathbf{x}_*$, we get a posterior mean (orange) & variance (grey region; $\pm 2\times$ standard deviation)



SUMMARY: BAYESIAN LM

- Bayesian perspective: entire distributions, rather than just point estimates, for $\boldsymbol{\theta}$
- From posterior distribution of $\boldsymbol{\theta}$ we can derive a predictive distribution for $y_* = \boldsymbol{\theta}^T \mathbf{x}_*$
- Online updates: after observing new data points, update posterior $\Rightarrow$ decreasing uncertainty