

Lecture 9: Linear Regression

Yi, Yung (이웅)

Mathematics for Machine Learning
<https://yung-web.github.io/home/courses/mathml.html>
 KAIST EE

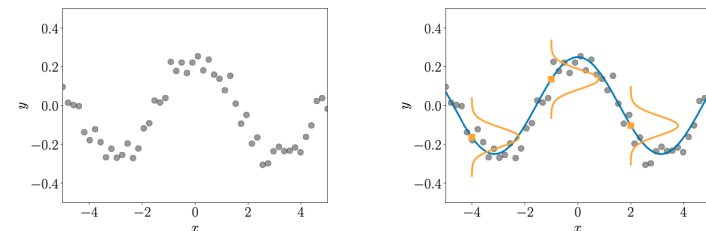
MONTH DAY, 2021

- Problem Formulation
- Parameter Estimation: ML
- Parameter Estimation: MAP
- Bayesian Linear Regression
- Maximum Likelihood as Orthogonal Projection

MONTH DAY, 2021 1 / 32

MONTH DAY, 2021 2 / 32

- Problem Formulation
- Parameter Estimation: ML
- Parameter Estimation: MAP
- Bayesian Linear Regression
- Maximum Likelihood as Orthogonal Projection



- For some input values x_n , we observe noisy function values $y_n = f(x_n) + \epsilon$
- Goal: infer the function f that generalizes well to function values at new inputs
- Applications: time-series analysis, control and robotics, image recognition, etc.

MONTH DAY, 2021 3 / 32

MONTH DAY, 2021 4 / 32

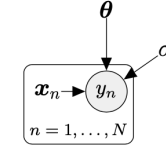
- Linear regression, Gaussian noise
- Notation for simplification (this is how the textbook uses)

$$p(y|\mathbf{x}) = p_{Y|\mathbf{X}}(y|\mathbf{x}), \quad Y \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{\text{simplifies}} \mathcal{N}(y | f(\mathbf{x}), \sigma^2)$$

- Likelihood: for $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathbb{R}$, $p(y | \mathbf{x}) = \mathcal{N}(y | f(\mathbf{x}), \sigma^2)$
- $y = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Linear regression with the parameter $\boldsymbol{\theta} \in \mathbb{R}^D$

$$p(y | \mathbf{x}) = \mathcal{N}(y | \mathbf{x}^T \boldsymbol{\theta}, \sigma^2) \iff y = \mathbf{x}^T \boldsymbol{\theta} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Prior with Gaussian noise: $p(y | \mathbf{x}) = \mathcal{N}(y | \mathbf{x}^T \boldsymbol{\theta}, \sigma^2)$



- Training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$

- Assuming iid of N data, the likelihood is factorized into:

$$p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta}) = \prod_{n=1}^N p(y_n | \mathbf{x}_n, \boldsymbol{\theta}) = \prod_{n=1}^N \mathcal{N}(y_n | \mathbf{x}_n^T \boldsymbol{\theta}, \sigma^2),$$

where $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and $\mathcal{Y} = \{y_1, \dots, y_N\}$

- ML and MAP

- Problem Formulation
- **Parameter Estimation: ML**
- Parameter Estimation: MAP
- Bayesian Linear Regression
- Maximum Likelihood as Orthogonal Projection

- $\boldsymbol{\theta}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta}) = \arg \min_{\boldsymbol{\theta}} (-\log p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta}))$

- For Gaussian noise with $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$ and $\mathbf{y} = [y_1, \dots, y_N]^T$,

$$\begin{aligned} -\log p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta}) &= -\log \prod_{n=1}^N p(y_n | \mathbf{x}_n, \boldsymbol{\theta}) = -\sum_{n=1}^N \log p(y_n | \mathbf{x}_n, \boldsymbol{\theta}) \\ &= \frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - \mathbf{x}_n^T \boldsymbol{\theta})^2 + \text{const} = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \text{const} \end{aligned}$$

Negative-log likelihood for $f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$:

$$-\log p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \text{const}$$

- For Gaussian noise with $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ and $\mathbf{y} = [y_1, \dots, y_n]^T$,

$$\theta_{\text{ML}} = \arg \min_{\theta} \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\theta\|^2, \quad L(\theta) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\theta\|^2$$
- In this special case of Gaussian noise, finding MLE is equivalent to finding θ that minimizes the empirical risk with squared loss function
 - Models as functions = Model as probabilistic models
- We find θ such that $\frac{dL}{d\theta} = 0$

$$\begin{aligned} \frac{dL}{d\theta} &= \frac{1}{2\sigma^2} (-2(\mathbf{y} - \mathbf{X}\theta)^T \mathbf{X}) = \frac{1}{\sigma^2} (-\mathbf{y}^T \mathbf{X} + \theta^T \mathbf{X}^T \mathbf{X}) = 0 \\ &\iff \theta_{\text{ML}}^T \mathbf{X}^T \mathbf{X} = \mathbf{y}^T \mathbf{X} \\ &\iff \theta_{\text{ML}}^T = \mathbf{y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (\mathbf{X}^T \mathbf{X} \text{ is positive definite if } \text{rk}(\mathbf{X}) = D) \\ &\iff \theta_{\text{ML}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

- Linear regression: Linear in the parameters
 - $\phi(\mathbf{x})^T \theta$ is also fine, where $\phi(\mathbf{x})$ can be non-linear (we will cover this later)
 - $\phi(\mathbf{x})$ are the features

- Linear regression with the parameter $\theta \in \mathbb{R}^K$, $\phi(\mathbf{x}) : \mathbb{R}^D \mapsto \mathbb{R}^K$:

$$p(y | \mathbf{x}) = \mathcal{N}(y | \phi(\mathbf{x})^T \theta, \sigma^2) \iff y = \phi(\mathbf{x})^T \theta + \epsilon = \sum_{k=0}^{K-1} \theta_k \phi_k(\mathbf{x}) + \epsilon$$

- Example. Polynomial regression.** For $x \in \mathbb{R}$ and $\theta \in \mathbb{R}^K$, we lift the original 1-D input into K -D feature space with monomials x^k :

$$\phi(x) = \begin{pmatrix} \phi_0(x) \\ \vdots \\ \phi_{K-1}(x) \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ x^{K-1} \end{pmatrix} \in \mathbb{R}^K \implies f(x) = \sum_{k=0}^{K-1} \theta_k x^k$$

- Now, for the entire training set $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$,

$$\Phi := \begin{pmatrix} \phi^T(\mathbf{x}_1) \\ \vdots \\ \phi^T(\mathbf{x}_N) \end{pmatrix} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \cdots & \phi_{K-1}(\mathbf{x}_1) \\ \vdots & \cdots & \vdots \\ \phi_0(\mathbf{x}_N) & \cdots & \phi_{K-1}(\mathbf{x}_N) \end{pmatrix} \in \mathbb{R}^{N \times K}, \quad \Phi_{ij} = \phi_j(\mathbf{x}_i), \quad \phi_j : \mathbb{R}^D \mapsto \mathbb{R}$$

- Negative log-likelihood: Similarly to the case of $\mathbf{y} = \mathbf{X}\theta$,

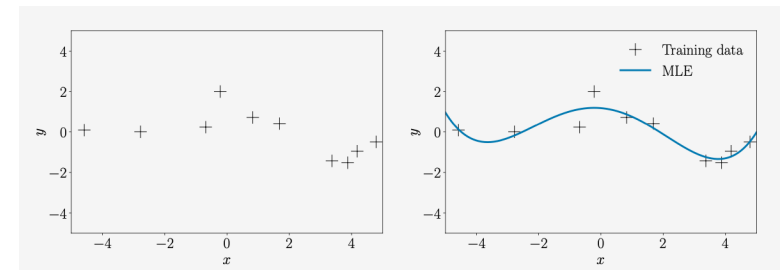
$$\circ p(\mathcal{Y} | \mathcal{X}, \theta) = \mathcal{N}(\mathbf{y} | \Phi\theta, \sigma^2 I)$$

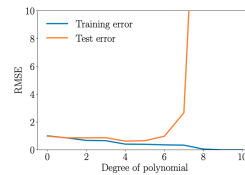
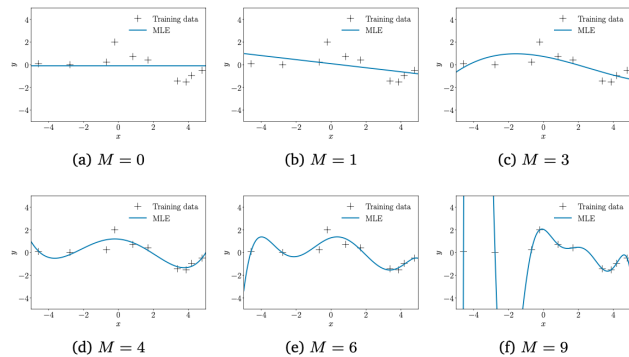
$$\circ \text{Negative-log likelihood for } f(\mathbf{x}) = \phi^T(\mathbf{x})\theta + \mathcal{N}(0, \sigma^2):$$

$$-\log p(\mathcal{Y} | \mathcal{X}, \theta) = \frac{1}{2\sigma^2} \|\mathbf{y} - \Phi\theta\|^2 + \text{const}$$

- MLE: $\theta_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$

- $N = 10$ data, where $x_n \sim \mathcal{U}[-5, 5]$ and $y_n = -\sin(x_n/5) + \cos(x_n) + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.2^2)$
- Fit with polynomial with degree 4 using ML





- Higher polynomial degree is better (training error always decreases)
- Test error increases after some polynomial degree

MONTH DAY, 2021 13 / 32

MONTH DAY, 2021 14 / 32

- Problem Formulation
- Parameter Estimation: ML
- **Parameter Estimation: MAP**
- Bayesian Linear Regression
- Maximum Likelihood as Orthogonal Projection

- MLE: prone to overfitting, where the magnitude of the parameters becomes large.
- a prior distribution $p(\theta)$ helps: what θ is plausible
- MAPE and Bayes' theorem

$$p(\theta | \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} | \mathcal{X}, \theta)p(\theta)}{p(\mathcal{Y} | \mathcal{X})} \Rightarrow \theta_{\text{MAP}} \in \arg \min_{\theta} \left(-\log p(\mathcal{Y} | \mathcal{X}, \theta) - \log p(\theta) \right)$$

- Gradient

$$-\frac{d \log p(\theta | \mathcal{X}, \mathcal{Y})}{d\theta} = -\frac{d \log p(\mathcal{Y} | \mathcal{X}, \theta)}{d\theta} - \frac{d \log p(\theta)}{d\theta}$$

- **Example.** A (conjugate) Gaussian prior $p(\theta) \sim \mathcal{N}(0, b^2 \mathbf{I})$
 - For Gaussian likelihood, Gaussian prior \Rightarrow Gaussian posterior
- Negative log-posterior

$$\begin{aligned} &\text{Negative-log posterior for } f(\mathbf{x}) = \phi^T(\mathbf{x})\theta + \mathcal{N}(0, \sigma^2) \text{ and } p(\theta) \sim \mathcal{N}(0, b^2 \mathbf{I}): \\ &-\log p(\theta | \mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2}(\mathbf{y} - \Phi\theta)^T(\mathbf{y} - \Phi\theta) + \frac{1}{2b^2}\theta^T\theta + \text{const} \end{aligned}$$

- Gradient
- $$-\frac{d \log p(\theta | \mathcal{X}, \mathcal{Y})}{d\theta} = \frac{1}{\sigma^2}(\theta^T \Phi^T \Phi - \mathbf{y}^T \Phi) + \frac{1}{b^2}\theta^T$$

MONTH DAY, 2021 15 / 32

MONTH DAY, 2021 16 / 32

- MAP vs. ML

$$\theta_{\text{MAP}} = \underbrace{\left(\Phi^T \Phi + \frac{\sigma^2}{b^2} I \right)}_{(*)}^{-1} \Phi^T \mathbf{y}, \quad \theta_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

- The term $\frac{\sigma^2}{b^2} I$
 - Ensures that $(*)$ is symmetric, strictly positive definite
 - Role of regularizer

- **Example.** A (conjugate) Gaussian prior $p(\theta) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$
- Negative log-posterior

Negative-log posterior for $f(\mathbf{x}) = \phi^T(\mathbf{x})\theta + \mathcal{N}(0, \sigma^2)$ and $p(\theta) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$:

$$-\log p(\theta|\mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2} (\mathbf{y} - \Phi\theta)^T (\mathbf{y} - \Phi\theta) + \frac{1}{2} (\theta - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\theta - \mathbf{m}_0) + \text{const}$$

- We will use this later for computing the parameter posterior distribution in Bayesian linear regression.

- **Explicit regularizer** in regularized least squares (RLS)

$$\|\mathbf{y} - \Phi\theta\|^2 + \lambda \|\theta\|^2$$

- **MAPE with Gaussian prior** $p(\theta) \sim \mathcal{N}(0, b^2 I)$
 - Negative log-Gaussian prior

$$-\log p(\theta) = \frac{1}{2b^2} \theta^T \theta + \text{const}$$

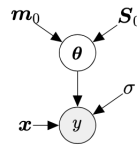
- $\lambda = 1/2b^2$ is the regularization term
- Not surprising that we have

$$\theta_{\text{RLS}} = \left(\Phi^T \Phi + \lambda I \right)^{-1} \Phi^T \mathbf{y}$$

- Problem Formulation
- Parameter Estimation: ML
- Parameter Estimation: MAP
- **Bayesian Linear Regression**
- Maximum Likelihood as Orthogonal Projection

- Earlier, ML and MAP. Now, **fully Bayesian**
- Model

$$\begin{array}{ll} \text{prior} & p(\theta) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0) \\ \text{likelihood} & p(y|\mathbf{x}, \theta) \sim \mathcal{N}(y | \phi^T(\mathbf{x})\theta, \sigma^2) \\ \text{joint} & p(y, \theta|\mathbf{x}) = p(y|\mathbf{x}, \theta)p(\theta) \end{array}$$



- Goal: For an input \mathbf{x}_* , we want to compute the following **posterior predictive distribution** of y_* :

$$p(y_*|\mathbf{x}_*, \mathcal{X}, \mathcal{Y}) = \int \overbrace{p(y_*|\mathbf{x}_*, \theta)}^{\text{likelihood}} \overbrace{p(\theta|\mathcal{X}, \mathcal{Y})}^{(*)} d\theta$$

- (*): parameter posterior distribution that needs to be computed

MONTH DAY, 2021 21 / 32

MONTH DAY, 2021 22 / 32

- Parameter posterior distribution

$$p(\theta | \mathcal{X}, \mathcal{Y}) = \mathcal{N}(\theta | \mathbf{m}_N, \mathbf{S}_N), \quad \text{where}$$

$$\mathbf{S}_N = (\mathbf{S}_0^{-1} + \sigma^2 \Phi^T \Phi)^{-1}, \quad \mathbf{m}_N = \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \sigma^{-2} \Phi^T \mathbf{y})$$

(Proof of Sketch)

- From the negative-log posterior for general Gaussian prior,

$$-\log p(\theta|\mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2} (\mathbf{y} - \Phi\theta)^T (\mathbf{y} - \Phi\theta) + \frac{1}{2} (\theta - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\theta - \mathbf{m}_0) + \text{const}$$

$$\begin{aligned} &= \frac{1}{2} \left(\sigma^{-2} \mathbf{y}^T \mathbf{y} - 2\sigma^{-2} \mathbf{y}^T \Phi \theta + \theta^T \sigma^{-2} \Phi^T \Phi \theta + \theta^T \mathbf{S}_0^{-1} \theta - 2\mathbf{m}_0^T \mathbf{S}_0^{-1} \theta + \mathbf{m}_0^T \mathbf{S}_0^{-1} \mathbf{m}_0 \right) \\ &= \frac{1}{2} \left(\theta^T (\sigma^{-2} \Phi^T \Phi + \mathbf{S}_0^{-1}) \theta - 2(\sigma^{-2} \Phi^T \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0)^T \theta \right) + \text{const} \end{aligned}$$

- **cyan color**: quadratic term, **orange color**: linear term
- $p(\theta|\mathcal{X}, \mathcal{Y}) \propto \exp(\text{quadratic in } \theta) \implies$ Gaussian distribution
- Assume that $p(\theta|\mathcal{X}, \mathcal{Y}) = \mathcal{N}(\theta|\mathbf{m}_N, \mathbf{S}_N)$, and find \mathbf{m}_N and \mathbf{S}_N .

$$\begin{aligned} -\log \mathcal{N}(\theta|\mathbf{m}_N, \mathbf{S}_N) &= \frac{1}{2} (\theta - \mathbf{m}_N)^T \mathbf{S}_N^{-1} (\theta - \mathbf{m}_N) + \text{const} \\ &= \frac{1}{2} \left(\theta^T \mathbf{S}_N^{-1} \theta - 2\mathbf{m}_N^T \mathbf{S}_N^{-1} \theta + \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N \right) + \text{const} \end{aligned}$$

Thus,

$$\mathbf{S}_N^{-1} = \sigma^{-2} \Phi^T \Phi + \mathbf{S}_0^{-1} \quad \text{and} \quad \mathbf{m}_N^T \mathbf{S}_N^{-1} = (\sigma^{-2} \Phi^T \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0^T)$$

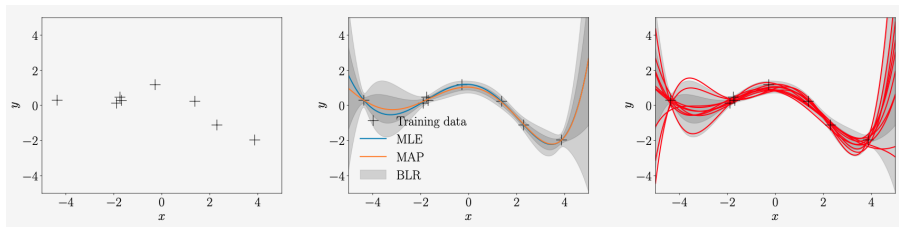
MONTH DAY, 2021 23 / 32

MONTH DAY, 2021 24 / 32

- Posterior predictive distribution

$$\begin{aligned} p(y_*|\mathbf{x}_*, \mathcal{X}, \mathcal{Y}) &= \int p(y_*|\mathbf{x}_*, \theta) p(\theta|\mathcal{X}, \mathcal{Y}) d\theta \\ &= \int \mathcal{N}(y_* | \phi^T(\mathbf{x}_*)\theta, \sigma^2) \mathcal{N}(\theta|\mathbf{m}_N, \mathbf{S}_N) d\theta \\ &= \mathcal{N}(y_* | \phi^T(\mathbf{x}_*) \mathbf{m}_N, \phi^T(\mathbf{x}_*) \mathbf{S}_N \phi(\mathbf{x}_*) + \sigma^2) \end{aligned}$$

- The mean $\phi^T(\mathbf{x}_*) \mathbf{m}_N$ coincides with the MAP estimate



- BLR: Bayesian Linear Regression

MONTH DAY, 2021 25 / 32

MONTH DAY, 2021 26 / 32

- Problem Formulation
- Parameter Estimation: ML
- Parameter Estimation: MAP
- Bayesian Linear Regression
- Maximum Likelihood as Orthogonal Projection

- For $f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\theta}_{\text{ML}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \frac{\mathbf{X}^T \mathbf{y}}{\mathbf{X}^T \mathbf{X}} \in \mathbb{R}$

$$\mathbf{X} \boldsymbol{\theta}_{\text{ML}} = \frac{\mathbf{X} \mathbf{X}^T}{\mathbf{X}^T \mathbf{X}} \mathbf{y}$$

- Orthogonal projection of \mathbf{y} onto the one-dimensional subspace spanned by \mathbf{X}

- For $f(\mathbf{x}) = \boldsymbol{\phi}^T(\mathbf{x}) \boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\theta}_{\text{ML}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{y} = \frac{\boldsymbol{\Phi}^T \mathbf{y}}{\boldsymbol{\Phi}^T \boldsymbol{\Phi}} \in \mathbb{R}$

$$\boldsymbol{\Phi} \boldsymbol{\theta}_{\text{ML}} = \frac{\boldsymbol{\Phi} \boldsymbol{\Phi}^T}{\boldsymbol{\Phi}^T \boldsymbol{\Phi}} \mathbf{y}$$

- Orthogonal projection of \mathbf{y} onto the K -dimensional subspace spanned by columns of $\boldsymbol{\Phi}$

MONTH DAY, 2021 27 / 32

MONTH DAY, 2021 28 / 32

- Linear regression for Gaussian likelihood and conjugate Gaussian priors. Nice analytical results and closed forms
- Other forms of likelihoods for other applications (e.g., classification)
- GLM (generalized linear model): $y = \sigma \circ f$ (σ : activation function)
 - No longer linear in θ
 - Logistic regression: $\sigma(f) = \frac{1}{1 + \exp(-f)} \in [0, 1]$ (interpreted as the probability of becoming 1)
 - Building blocks of (deep) feedforward neural nets
 - $\mathbf{y} = \sigma(\mathbf{Ax} + \mathbf{b})$. \mathbf{A} : weight matrix, \mathbf{b} : bias vector
 - K -layer deep neural nets: $\mathbf{x}_{k+1} = f_k(\mathbf{x}_k)$, $f_k(\mathbf{x}_k) = \sigma_k(\mathbf{A}_k \mathbf{x}_k + \mathbf{b}_k)$

- Gaussian process
 - A distribution over parameters \rightarrow a distribution over functions
 - Gaussian process: distribution over functions without detouring via parameters
 - Closely related to BLR and support vector regression, also interpreted as Bayesian neural network with a single hidden layer and the infinite number of units
- Gaussian likelihood, but non-Gaussian prior
 - When $N \ll D$ (small training data)
 - Prior that enforces sparsity, e.g., Laplace prior
 - A linear regression with the Laplace prior = linear regression with LASSO (L1 regularization)

Questions?

1)