Applications

-- (Bayesian) Regression where we have joint distribution over multiple predictions

-- Optimizing functions for which gradients are not available

E.g. Hyper-parameter optimization of deep models

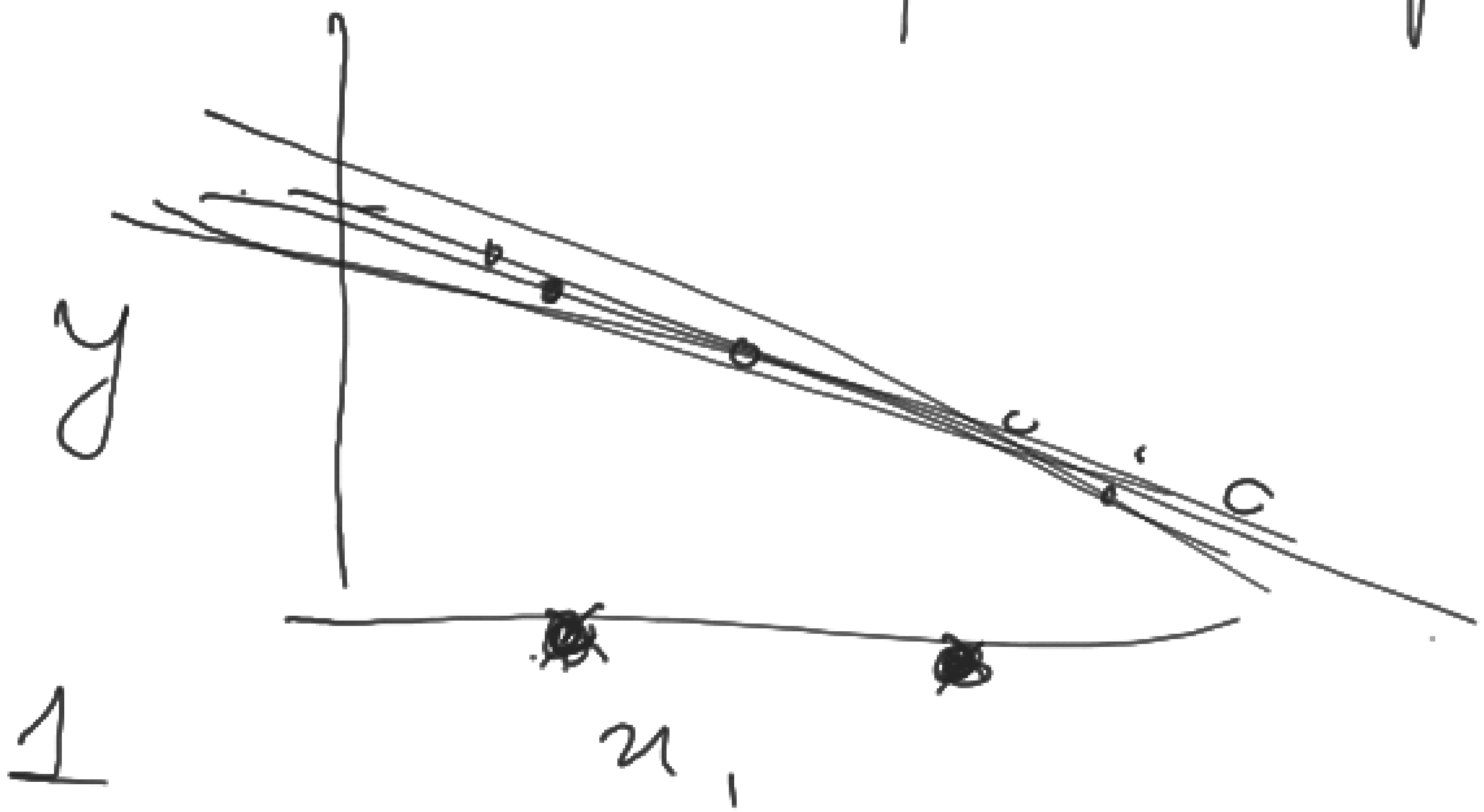
Normal Regression

$x \in \mathbb{R}^d$

$$f(x) = w_1 x_1 + w_2 x_2 + \ldots w_d x_d + b$$

$$y \sim \mathcal{N}(f(x), \sigma^2)$$

$w_1 \ldots w_d, b$ are parameters



$d = 1$

Gaussian Processes:

$x \in \mathbb{R}^d$

$f(\ )$

$x^1, x^2, x^3 \dots x^N$

$\underline{f}(x^1) \quad \underline{f}(x^2) \quad \dots \quad \underline{f(x^N)} \quad \sigma^2$

$$P\left(\begin{bmatrix} f(x^1) \\ f(x^2) \\ \vdots \\ f(x^N) \end{bmatrix}\right) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} K(x^1, x^1) & & \\ & K(x^i, x^j) & \\ & & k(x^N, x^N) \end{bmatrix}\right)$$

$\mu_{N\times 1} \quad \Sigma_{N\times N}$

$\Sigma_{ij} = K(x^i, x^j)$

$$P\left(\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}\right) \sim \mathcal{N}(\mu ; \Sigma) =$$

$$\mathcal{N}\left(\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{Y}, \mu, \Sigma_{N\times N}\right) = \frac{1 e^{-\frac{(Y-\mu)^T \Sigma^{-1} (Y-\mu)}{2}}}{\left(2\pi|\Sigma|\right)^{k/2}}$$

<u>Kernel function:</u>

RBF kernel: $e^{-\frac{\|x^i - x^j\|^2}{l}}$ l ← length parameter

$x^i, x^j \in \mathbb{R}^d$

Properties of multivariate Gaussians.

① $Y_A \sim \mathcal{N}(\mu_A ; \Sigma_{AA})$. $Y_B \sim \mathcal{N}(\mu_B , \Sigma_{BB})$ $Y_A \perp\!\!\!\perp Y_B$

$Y_A + Y_B \sim \mathcal{N}(\mu_A + \mu_B ; \Sigma_{AA} + \Sigma_{BB})$

$$Y = \begin{bmatrix} Y_A \\ Y_B \end{bmatrix} \sim N(\mu ; \Sigma) = N\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} ; \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right)$$

$$Y = \begin{array}{c|c} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \hline \\ \\ y_N \end{bmatrix} & \begin{array}{c} Y_A \\ \\ \\ \\ Y_B \end{array} \end{array}$$

$$P(Y_A \mid Y_B = O_B) = N(\mu_{A|B} ; \Sigma_{A|B})$$

$$\mu_{A|B} = \mu_A + \Sigma_{AB} \Sigma_{BB}^{-1} (O_B - \mu_B)$$

$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}$$

Posterior distribution of the function

$$Y \equiv \left[\begin{array}{c} y_1 \\ \vdots \\ y_N \\ \hline y^* \end{array}\right] \quad \vec{Y} \qquad \left[\begin{array}{c} x^1 \\ \vdots \\ x^N \\ \hline x^* \end{array}\right] \quad \mathbf{X}$$

$$P\left(y^* \,\middle|\, \underbrace{\begin{bmatrix} f(x^1) = y_1 \\ f(x^2) = y_2 \\ \vdots \\ f(x^N) = y_N \end{bmatrix}}_{\text{training data}}, \mathbf{x}^* \right)$$

$\mathcal{N}(\mu_*, \sigma^*)$ where

$$\mu_* = 0 + k(\mathbf{x}^*, \mathbf{X})\left[K(\mathbf{X}, \mathbf{X})\right]^{-1} \vec{Y}$$

$$\sigma_*^2 = k(x^*, x^*) - k(x^*, \mathbf{X})\left[K(\mathbf{X}, \mathbf{X})\right]^{-1} k(\mathbf{X}, x^*)$$

Let's say we want to get distribution over

$$\left.\begin{matrix} x^{N+1} \\ \vdots \\ \vdots \\ x^{N+M} \end{matrix}\right\} X^{*} \qquad P\left(\underset{M\times 1}{Y^{*}} \,\middle|\, \begin{matrix} f(x^1)=y_1 \\ \vdots \\ f(x^N)=y_N \end{matrix}\right) \sim N\left(\underset{M\times 1}{\mu^{*}},\ \underset{M\times M}{\Sigma^{*}}\right)$$

$$\mu^{*} = 0 + K(X^{*}, X)\left[K(X, X)\right]^{-1} Y$$

$$\Sigma^{*} = K(X^{*}, X^{*}) - K(X^{*}, X)\left[K(X, X)\right]^{-1} K(X, X^{*})$$

Hyper-parameter optimization of an expensive deep model M

$\underline{X}$ ≡ space of hyper-parameters.

eg. # layers, learning rate, # dimensions of hidden unit.

etc. vocab. size.

$k(x^i, x^j)$ ←

$\boxed{f(x)}$ ≡ validation loss or error of M with hyper-param x.

Goal: find the hyper-param for which $f(x)$ is minimum.

$f(x)$ is not differentiable.

For more on use of GPs for hyper-parameter optimization see: https://www.cs.cornell.edu/courses/cs4787/2019sp/notes/lecture16.pdf