PRML 3.2

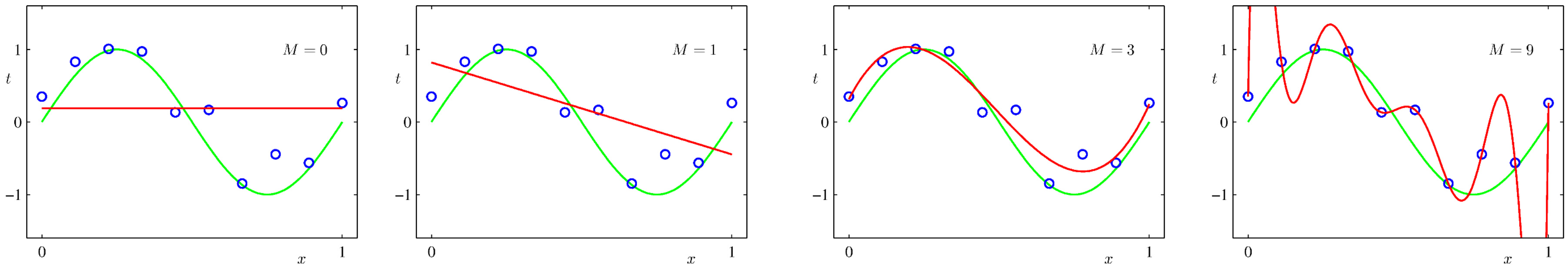
11/13/2019

Sakai Kazunori

3.2 バイアス-バリアンス分解

モチベーション

解きたい問題に合わせて
適切な複雑さのモデルを選択したい



頻度主義においてはそれをどう果たすのか？

キーワード

バイアス、バリアンス

一言で

バイアス

モデル近似誤差。

バリアンス

統計的推定誤差。

予測分布の作り方

- **最尤推測**
$$= p(\mathbf{x}|\mathbf{w}_{\mathrm{ML}})$$

- **平均プラグイン推測**
$$= p(\mathbf{x}|\mathbb{E}_{\mathbf{w}}[\mathbf{w}])$$

- **MAP推測**
$$= p(\mathbf{x}|\mathbf{w}_{\mathrm{MAP}})$$

- **ベイズ推測**
$$= \mathbb{E}_{\mathbf{w}}[p(\mathbf{x}|\mathbf{w})]$$

※cited from sdm-190529-information_criteria-enomoto-sakai-part1.pptx

予測分布の作り方

- **最尤推測**
  $= p(\mathbf{x}|\mathbf{w}_{\mathrm{ML}})$
- **平均プラグイン推測**
  $= p(\mathbf{x}|\mathbb{E}_{\mathbf{w}}[\mathbf{w}])$
- **MAP推測**
  $= p(\mathbf{x}|\mathbf{w}_{\mathrm{MAP}})$
- **ベイズ推測**
  $= \mathbb{E}_{\mathbf{w}}[p(\mathbf{x}|\mathbf{w})]$

回帰問題（のアプローチの１つ）

Given

入力: $\mathbf{x}$

教師: $t = g(\mathbf{x}) + \text{noise}$

モデル: $y = f(\mathbf{x}, \mathbf{w}) + \text{noise}$

任意の誤差関数: $\mathcal{L}(\mathbf{x}, \mathbf{w}, t)$

$$y(\mathbf{x}; \mathcal{D}) = f(\mathbf{x}, \mathbf{w}^*) + \text{noise}$$

$\mathbf{w}_{\text{ML}}$ とか、$\mathbf{w}_{\text{MAP}}$ とか、

学習の目標

$$\text{教師:} \quad t = g(\mathbf{x}) + \text{noise}$$

$$\text{学習後のモデル:} \quad y(\mathbf{x}; \mathcal{D}) = f(\mathbf{x}, \mathbf{w}^*) + \text{noise}$$

訓練誤差

$$\frac{1}{|\mathcal{D}|} \sum \mathcal{L}(\mathbf{x}, \mathbf{w}^*, t)$$

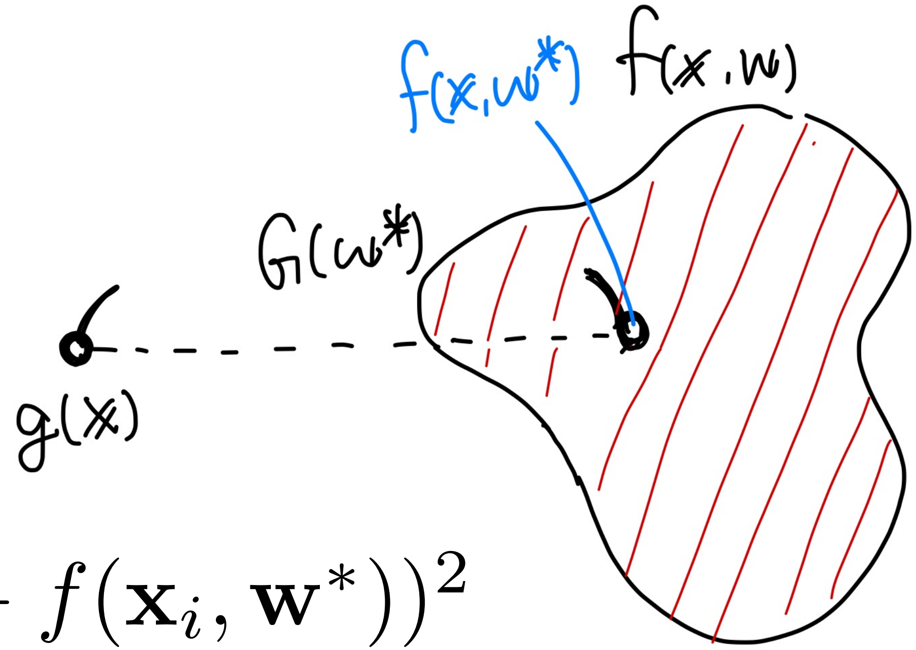
汎化誤差

$g(\mathbf{x})$ と $f(\mathbf{x}, \mathbf{w}^*)$ の差。

Lを二乗誤差関数とすれば？

訓練誤差と汎化誤差

二乗誤差を最小にするパラメータを $\mathbf{w}^*$ とする。



訓練誤差

$$T(\mathbf{w}^*) = \frac{1}{|\mathcal{D}|}\sum_i \left(t_i - f(\mathbf{x}_i, \mathbf{w}^*)\right)^2$$

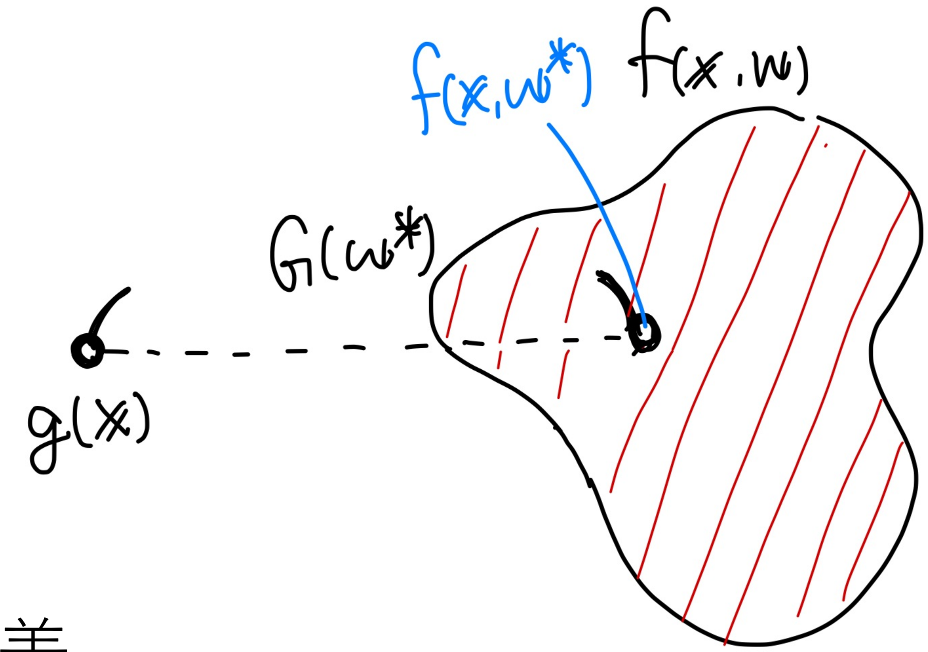
汎化誤差

$$G(\mathbf{w}^*) = \iint \left(t - f(\mathbf{x}, \mathbf{w}^*)\right)^2 p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t$$

汎化誤差の要因

バイアス モデル近似誤差。

どんなにパラメータを工夫しても
$f(\mathbf{x}, \mathbf{w})$ は $g(\mathbf{x})$ になれない。



バリアンス 統計的推定誤差。

データにはノイズが含まれており、
使う**データセットに応じてパラメータの推定がばらつく**。

実際に数式を分解してみる

回帰の期待損失関数 (1.5.5復習)

誤差関数: $\mathcal{L}(t, y(\mathbf{x}))$

期待損失(汎化誤差): $\mathbb{E}\left[\mathcal{L}\right] = \iint \mathcal{L}(t, y(\mathbf{x}))p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t$

$$h(\mathbf{x}) = \arg\min_{y} \mathbb{E}\left[\mathcal{L}\right]$$

神のみぞ知る

$\mathcal{L}(t, y(\mathbf{x}))$ が二乗誤差関数の場合、 $h(\mathbf{x}) = \mathbb{E}_t\left[t|\mathbf{x}\right]$

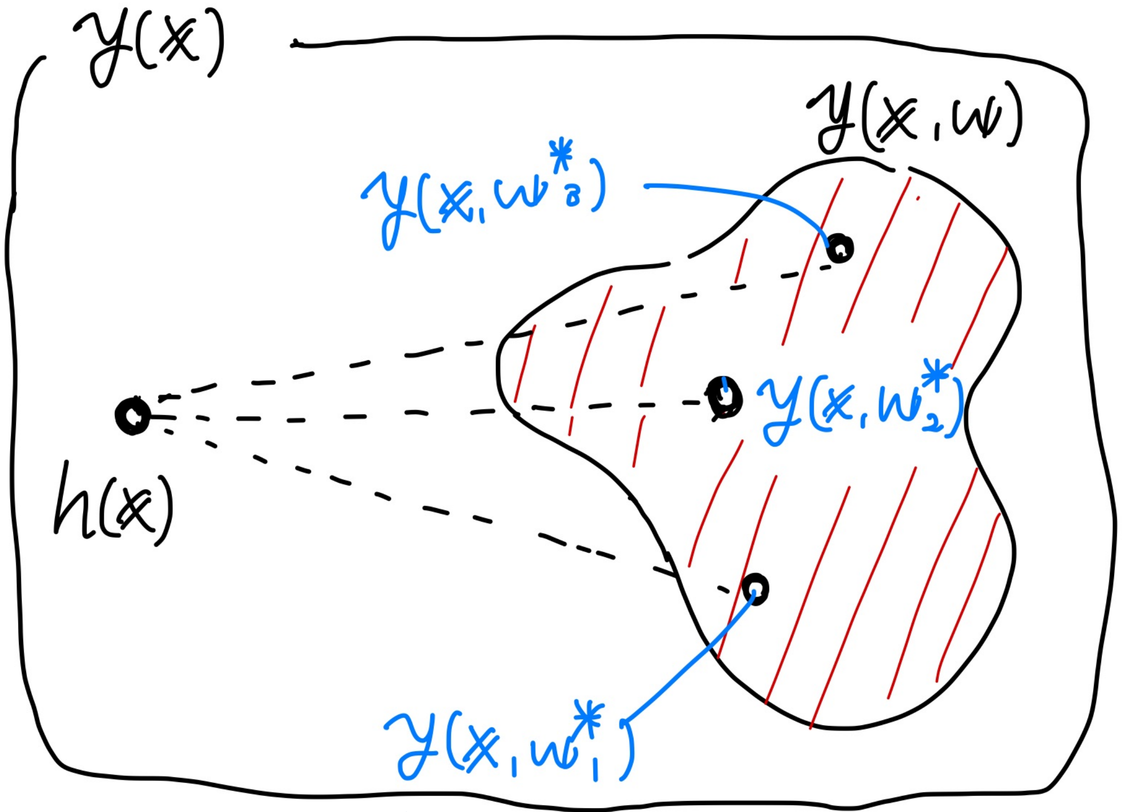
例えば

$y(\mathbf{x})$ を求める問題を、

$y(\mathbf{x}, \mathbf{w})$ のパラメータを $\mathbf{w}$ 求める問題に。

性能評価

データにはノイズが含まれるため、推定値はデータセットによってぶれる



$$\mathbf{w}_1^* = \arg\min_{\mathbf{w}} \sum_{\mathcal{D}_1} \mathcal{L}$$

$$\mathbf{w}_2^* = \arg\min_{\mathbf{w}} \sum_{\mathcal{D}_2} \mathcal{L}$$

$$\vdots$$

データセットについて平均して評価する。

期待二乗損失(汎化誤差)の期待値の分解

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}\left[\mathcal{L}\right]\right] = \mathbb{E}_{\mathcal{D}}\left[G\right]$$

$$= \int \left\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\right\}^2 p(\mathbf{x})\mathrm{d}\mathbf{x}$$

(バイアス)²

$$+ \int \mathbb{E}_{\mathcal{D}}\left[\left\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\right\}^2\right] p(\mathbf{x})\mathrm{d}\mathbf{x}$$

バリアンス

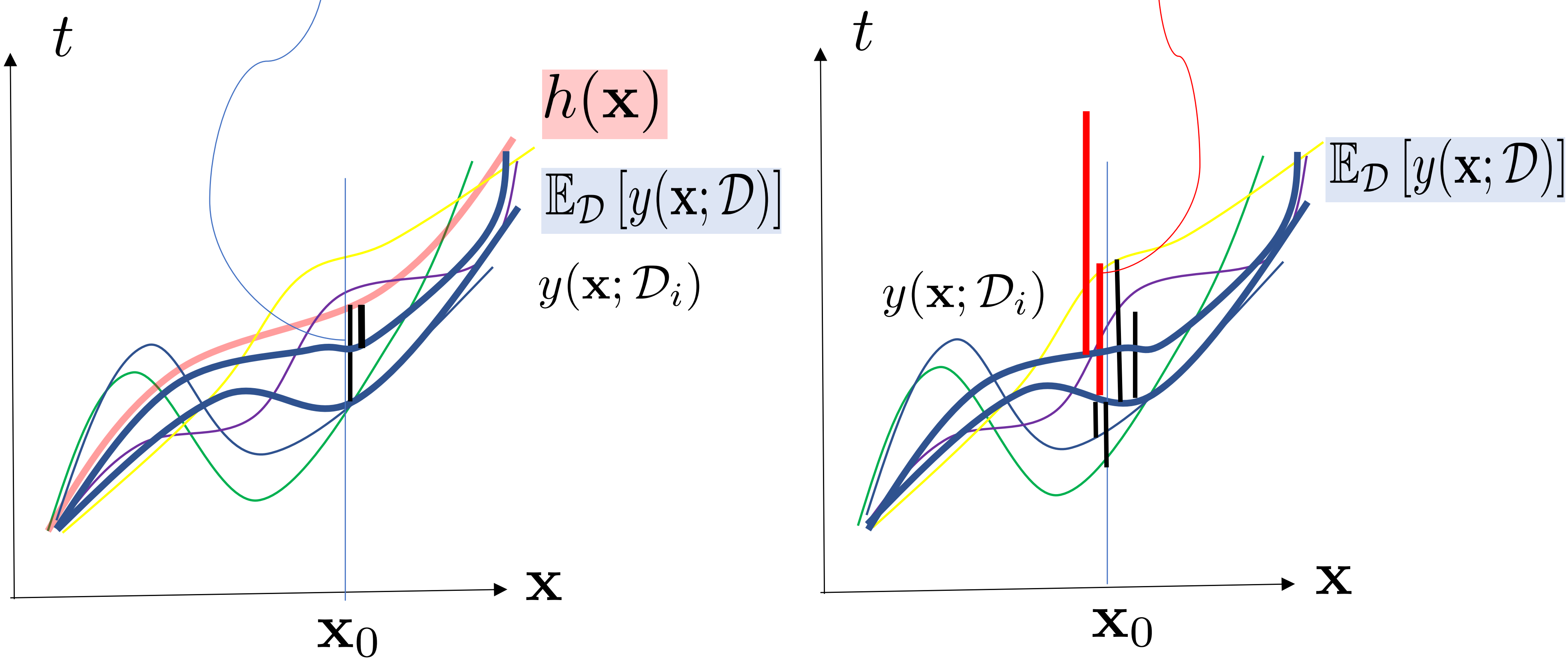
$$+ \iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t$$

ノイズ

バイアス-バリアンスの意味

(バイアス)^2 バリアンス

$$\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}_0;\mathcal{D})]-h(\mathbf{x}_0)\}^2+\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x}_0;\mathcal{D})-\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}_0;\mathcal{D})]\}^2\right]$$



トレードオフの関係

※cited from PRML_3.2-3.3.2_Enomoto.pptx

モデルの複雑さとバイアス-バリアンス

$$h(x) = \sin(2\pi x)$$

$$\mathcal{D}^{(l)}(l = 1, \ldots, L; L = 100)$$

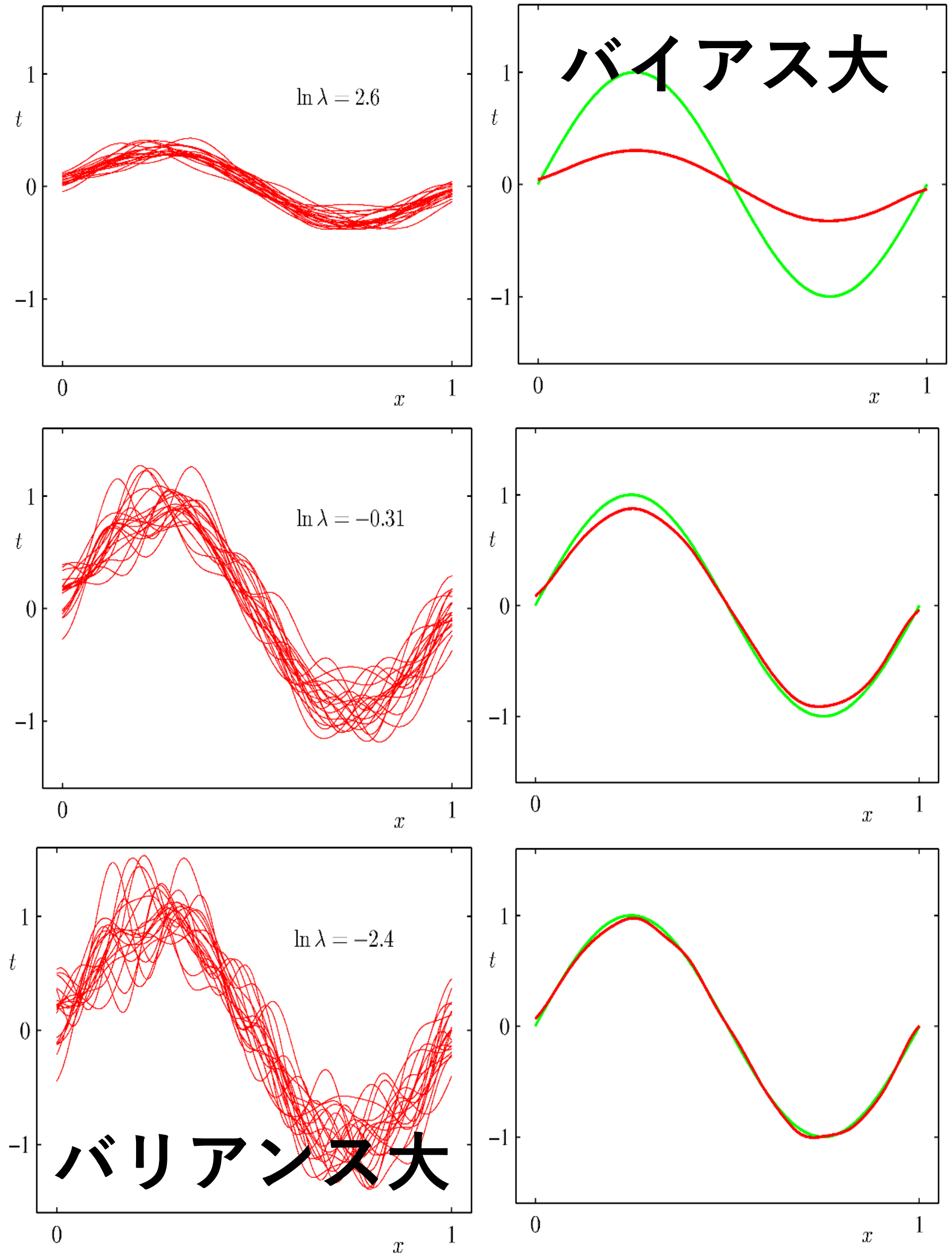
$$\mathcal{D} = \{x_n, t_n\}_{n=1}^{N=25}$$

$$\frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\left(\mathbf{x}_n\right)\right\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

100個の回帰を得る

平均をとることは有用かも？

単なる平均と
ベイズ推測の違いは？



予測分布の作り方

- **最尤推測**

  $= p(\mathbf{x}|\mathbf{w}_{\mathrm{ML}})$

- **平均プラグイン推測**

  $= p(\mathbf{x}|\mathbb{E}_{\mathbf{w}}[\mathbf{w}])$

- **MAP推測**

  $= p(\mathbf{x}|\mathbf{w}_{\mathrm{MAP}})$

- **ベイズ推測**

  $= \mathbb{E}_{\mathbf{w}}[p(\mathbf{x}|\mathbf{w})]$

※cited from sdm-190529-information_criteria-enomoto-sakai-part1.pptx

トレードオフの考察

$$(\text{ bias })^2 = \frac{1}{N} \sum_{n=1}^{N} \{ \bar{y}(x_n) - h(x_n) \}^2$$

$$\text{variance } = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{L} \sum_{l=1}^{L} \left\{ y^{(l)}(x_n) - \bar{y}(x_n) \right\}^2$$

$$\bar{y}(x) = \frac{1}{L} \sum_{l=1}^{L} y^{(l)}(x) \approx \mathbb{E}_{\mathcal{D}} [y(x; \mathcal{D})]$$

データセットを複数用意することで
モデルの複雑さについて考察ができた。