Module 1: Least Squares

Localizationの定義
the method by which we determine the position and orientation of a vehicle within the world.

State Estimationの定義
the process of computing a physical quantity (that changes over time) like a position from a set of measurements.

Parameter Estimationの定義
a parameter is constant over time.
例えば、position and orientation are states of a moving vehicle, while the resistance of a particular resistor in the electrical sub-system of a vehicle would be a parameter.

内容
- Ordinary and weighted least squares
- Recursive least squares
- Link between Maximum likelihood and the method of least squares

Lesson 1 (Part 1): Squared Error Criterion and the Method of Least Squares

単語
- Ceres: the smallest dwarf planet in the solar system, located in the asteroid belt. It has a diameter of 930 kilometers. ケレス（準惑星、じゅんわくせい, dwarf planet）準惑星：太陽の周囲を公転する惑星以外の天体のうち、それ自身の重力によって球形になれるだけの質量を有するもの。
- Singular: a singular matrix is a square matrix which is not invertible. Alternatively, a matrix is singular if and only if it has a determinant of 0.

内容
- Describe the least error criterion and how it's used in parameter estimation.
- Derive the normal equations for least squares parameter estimation.

Giuseppe Piazzi and Ceresの発見
- Piazzi made 24 telescope observations of this new object over 40 days before it was lost in the glare of the sun.
  - Since Ceres is only about 900km in diameter, finding it again was extremely challenging.
  - This meant that other astronomers could not confirm Piazzi's discovery.
- Carl Friedrich Gauss used a method of least squares to accurately estimate Ceres orbital parameters based on Piazzi's published measurements.

Color-coded carbon film resistor
- the resistor has a gold band, which indicates that it can vary by as much as 5%.

Minimizing the Squared Error Criterion
- $$\begin{aligned} \mathscr{L}_{LS}(x) = e_1^2 + e_2^2 + e_3^2 + e_4^2 &= e^T e \\ &= (y - Hx)^T (y - Hx) \\ &= y^T y - x^T H^T y - y^T H x + x^T H^T H x \end{aligned}$$
- To minimize this, we can compute the partial derivative of the error function with respect to the unknown x and set the derivative to 0: $\frac{\partial \mathscr{L}}{\partial x}|_{x=\hat{x}} = -y^T H - y^T H + 2\hat{x}^T H^T H = 0$

- つまり　$\hat{x}_{LS} = (H^T H)^{-1} H^T y$
- We will only be able to solve for $\hat{x}$ if $(H^T H)^{-1}$ exists.
  - $H^T H$ is not singular.
- If we have m measurements and n unknown parameters: $H \in \mathbb{R}^{m \times n}$, $H^T H \in \mathbb{R}^{n \times n}$
  - This means that $(H^T H)^{-1}$ exists only if there are at least as many measurements as there are unknown parameters: $m \geq n$. measurementsの数は最低parametersの数、これは当たり前だ。
  - This will usually not be a problem. In fact, often face the challenge of dealing with too many measurements.

Method of Least Squares | Assumptions
- Measurement model, $y = x + v$, is linear.
  - often broken in complex systems.
- Measurements are equally weighted.
  - did not suspect that some have more noise than others.

Lesson 1 (Part 2): Squared Error Criterion and the Method of Least Squares

内容
- Derive the weighted least squares criterion given varying measurement noise variance.

Method of Weighted Least Squares
- Suppose we take measurements with multiple multimeters, some of which are better than others.
- Consider the general linear measurement model for m measurements and n unknowns:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \mathbf{H} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

  - $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}$
- In regular least squares, we implicitly assumed that each noise term was of equal variance:

$$\mathbb{E}[v_i^2] = \sigma^2, \mathbf{R} = \mathbb{E}[\mathbf{v}\mathbf{v}^T] = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$$

  - One way to interpret the ordinary method of least squares is to say that we are implicitly assuming that each noise term v_i is an independent random variable across measurements and has an equal variance or standard deviation. つまりI.I.D.条件。
- If we assume each noise term is independent, but of different variance. $\mathbb{E}[v_i^2] = \sigma_i^2$,

$$\mathbf{R} = \mathbb{E}[\mathbf{v}\mathbf{v}^T] = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_m^2 \end{bmatrix}$$

- Then we can define a weighted least squares criterion as:

$$\mathscr{L}_{WLS}(x) = e^T R^{-1} e$$
$$= \frac{e_1^2}{\sigma_1^2} + \frac{e_2^2}{\sigma_2^2} + \ldots + \frac{e_m^2}{\sigma_m^2}$$

where $\begin{bmatrix} e_1 \\ \vdots \\ e_m \end{bmatrix} = \mathbf{e} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} - \mathbf{H} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$

- Each squared error term is now weighted by the inverse of the variance associated with the corresponding measurement.
- In other words, the lower the variance of the noise, the more strongly its associated error term will be weighted in the loss function.
- We care more about errors which come from low noise measurements since those should tell us a lot about the true values of our unknown parameters.

Minimizing the Weighted Least Squares Criterion

- Expanding our new criterion: $$\mathscr{L}_{WLS}(x) = e^T R^{-1} e = (y - Hx)^T R^{-1}(y - Hx)$$
- Minimize it as before, but accounting for the new weighting term:
$$\frac{\partial \mathscr{L}}{\partial x}\Big|_{x=\hat{x}} = -y^T R^{-1} H - y^T R^{-1} H + 2\hat{x}^T H^T R^{-1} H = 0$$
  - The weighted normal equations: $H^T R^{-1} H \hat{x}_{WLS} = H^T R^{-1} y$
- 例:

	Resistance Measurements (Ohms)	
#	Multimeter A ($\sigma = 20$ ohms)	Multimeter B ($\sigma = 2$ ohms)
1	1068	
2	988	
3		1002
4		996

$$\hat{x}_{WLS} = (H^T R^{-1} H)^{-1} H^T R^{-1} y$$
$$= \left([1111] \begin{bmatrix} 400 & & & \\ & 400 & & \\ & & 4 & \\ & & & 4 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right)^{-1} [1111] \begin{bmatrix} 400 & & & \\ & 400 & & \\ & & 4 & \\ & & & 4 \end{bmatrix}^{-1} \begin{bmatrix} 1068 \\ 988 \\ 1002 \\ 996 \end{bmatrix}$$
$$= \frac{1}{1/400 + 1/400 + 1/4 + 1/4} \left(\frac{1068}{400} + \frac{988}{400} + \frac{1002}{4} + \frac{996}{4} \right)$$
$$= 999.3$$

- It's important to be comfortable working with different measurement variances and also with measurements that are sometimes correlated.
- A self-driving car will have a number of different and complex sensors on board and we need to make sure that we model our error sources correctly.
  - Accurate noise modeling is crucial to utilize various sensors effectively.

- Measurements can come from sensors that have different noisy characteristics.
- Weighted least squares lets us weight each measurement according to noise variance.

Lesson 2: Recursive Least Squares

単語

- Trace: In linear algebra, the trace (often abbreviated to tr) of a square matrix A is defined to be the sum of elements on the main diagonal (from the upper left to the lower right) of A.

compute least squares on the fly.

内容

- Extend the (batch) least squares formulation to a recursive one.

- Use this method to compute a 'running estimate' of the least squares solution as measurements stream in.

Linear Recursive Estimator

- an optimal estimate, $\hat{x}_{k-1}$, of unknown parameters at time $k-1$
- a new measurement at time k: $y_k = H_k x + v_k$
- Goal: compute $\hat{x}_k$ as a function of y_k and $\hat{x}_{k-1}$.
- linear recursive update: $\hat{x}_k = \hat{x}_{k-1} + K_k(y_k - H_k \hat{x}_{k-1})$.
  - K: estimator gain matrix
  - the term in brackets: innovation
    - Innovation quantifies how well current measurement matches previous best estimate.

K_kの計算

- Compute it by minimizing a similar least squares criterion, but this time use a probabilistic formulation.
- Wish to minimize the expected value of the sum of squared errors of our current estimate at time step k: $\mathscr{L}_{RLS} = \mathbb{E}[(x_k - \hat{x}_k)^2] = \sigma_k^2$
- If n unknown parameters at time step k, generalize to:
  $\mathscr{L}_{RLS} = \mathbb{E}[(x_{1k} - \hat{x}_{1k})^2 + \ldots + (x_{nk} - \hat{x}_{nk})^2] = Trace(P_k)$
  - P_k: estimator covariance
- Using linear recursive formulation, covariance can be expressed as a function of K_k: （証明略）
  $P_k = (1 - K_k H_k) P_{k-1} (1 - K_k H_k)^T + K_k R_k K_k^T$
- Minimized (through matrix calculus) when: $K_k = P_{k-1} H_k^T (H_k P_{k-1} H_k^T + R_k)^{-1}$ （証明略）
- With this expression, expression for P_k can also be simplified（R_kがなくなる、R_kはweight matrix）：$P_k = P_{k-1} - K_k H_k P_{k-1} = (1 - K_k H_k) P_{k-1}$
  - Covariance shrinks with each measurement.
  - The larger the gain matrix K_k, the smaller the new estimator covariance will be.
  - Intuitively, think this gain matrix as balancing the information from prior estimate and the information from new measurement.

Recursive Least Squares | Algorithms

1. Initialize the estimator: $\hat{x}_0 = \mathbb{E}[x]$, $P_0 = \mathbb{E}[(x - \hat{x}_0)(x - \hat{x}_0)^T]$
   1. This initial guess could come from the first measurement and the covariance could come from technical specifications (?).
2. Set up the measurement model, defining the Jacobian and the measurement covariance matrix: $y_k = H_k x + v_k$
3. Update the estimate $\hat{x}_k$ and the covariance P_k using:
$$K_k = P_{k-1} H_k^T (H_k P_{k-1} H_k^T + R_k)^{-1}$$
$$\hat{x}_k = \hat{x}_{k-1} + K_k (y_k - H_k \hat{x}_{k-1})$$
$$P_k = (1 - K_k H_k) P_{k-1}$$

Recursive Least SquaresのImportance

1. Minimize computational effort in estimation process.
2. Forms the update step of the linear Kalman filter.

- RLS produces a 'running estimate' of parameter(s) for a stream of measurements.
- RLS is a linear recursive estimator that minimizes the (co)variance of the parameter(s) at the current time.

Lesson 3: Least Squares and the Method of Maximum Likelihood

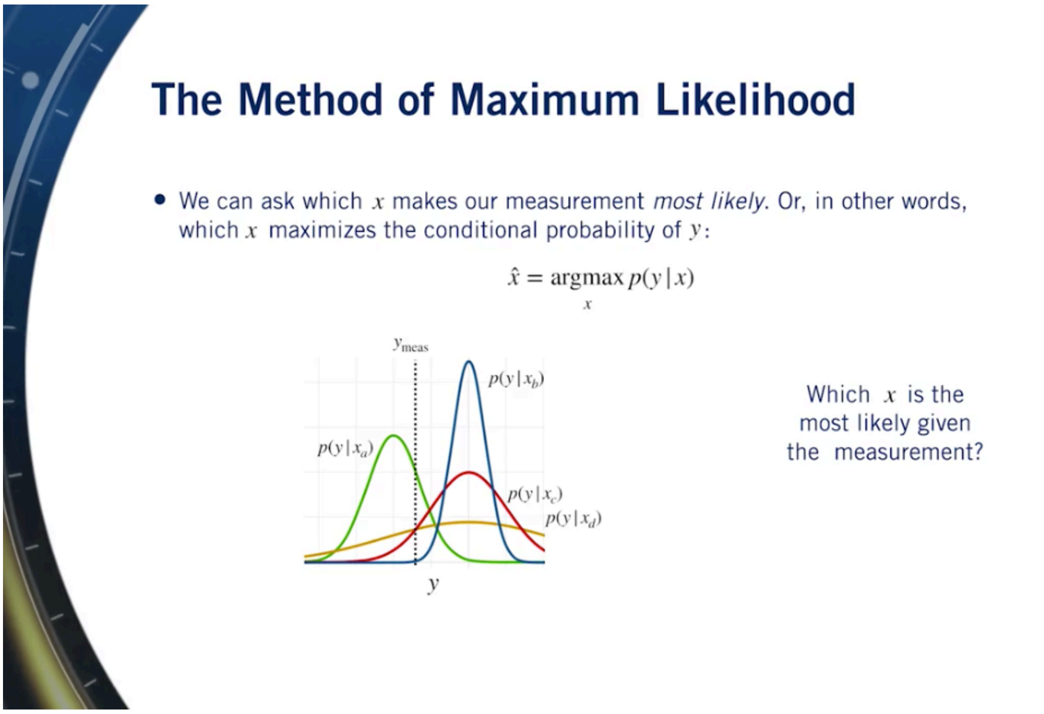
これは既にDeep Learningの中に読んだ。理解していた。

内容

- State the connection between the method of least squares and maximum likelihood with Gaussian random variables.

Why Squared Errors?

1. 計算簡単：If the measurement model is linear, minimizing the squared error criterion amounts to solving a linear system of equations.
2. Probability and a deep connection between least squares and maximum likelihood estimators under the assumption of Gaussian noise.



The Method of Maximum Likelihood（既に分かっていた）

- Ask which x makes measurement most likely. Or, in other words, which x maximizes the conditional probability of y: $\hat{x} = \underset{x}{argmax}\, p(y|x)$

Measurement Model

- Simple measurement model: $y = x + v$
- Convert this to a conditional probability on measurement, by assuming some probability density for v. For example, if $v \sim \mathcal{N}(0,\sigma^2)$
- Then, $p(y|x) = \mathcal{N}(x, \sigma^2)$
  - The unknown parameter x becomes the mean of this density, and the variance is noise variance.

$$p(y|x) = \mathcal{N}(y; x, \sigma^2)$$

- Using probability density function of a Gaussian: $$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-x)^2}{2\sigma^2}}$$
- If multiple independent measurements, then:
$$p(y|x) \propto \mathcal{N}(y_1; x, \sigma^2)\mathcal{N}(y_2; x, \sigma^2) \times \ldots \times \mathcal{N}(y_m; x, \sigma^2)$$
$$= \frac{1}{\sqrt{(2\pi)^m \sigma^{2m}}} exp\left(\frac{-\sum_{i=1}^{m}(y_i - x)^2}{2\sigma^2}\right)$$
- The maximal likelihood estimate (MLE) is given by: $$\hat{x}_{MLE} = \underset{x}{argmax}\, p(y|x)$$ $$= \underset{x}{argmax}\, logp(y|x)$$
  - The logarithm is monotonically increasing.
- Resulting in: $logp(y|x) = -\frac{1}{2R}((y_1 - x)^2 + \ldots + (y_m - x)^2) + C$
- Since $\underset{z}{argmax} f(z) = \underset{z}{argmin}(-f(z))$
- The maximum likelihood problem can therefore be written as
$$\hat{x}_{MLE} = \underset{x}{argmin} - logp(y|x)$$
$$= \underset{x}{argmin} \frac{1}{2\sigma^2}((y_1 - x)^2 + \ldots + (y_m - x)^2)$$
- Finally, if each measurement has a different variance,
$$\hat{x}_{MLE} = \underset{x}{argmin} \frac{1}{2}\left(\frac{(y_1 - x)^2}{\sigma_1^2} + \ldots + \frac{(y_m - x)^2}{\sigma_m^2}\right)$$
- In both cases, $\hat{x}_{MLE} = \hat{x}_{LS} = \underset{x}{argmin}\, \mathcal{L}_{LS}(x) = \underset{x}{argmax}\, \mathcal{L}_{MLE}(x)$

The Central Limit Theorem
- In realistic systems like self-driving cars, there are many sources of ‘noise’
- Central Limit Theorem: When independent random variables are added, their normalized sum tends towards a normal distribution.
  - model system probabilistically and yet maintain simplicity in calculations.

Least Squares | Some Caveats
- ‘Poor’ measurements (e.g. outliers) have a significant effect on the method of least squares.
  - Outliers might result from people walking in the middle of a Lidar scan, or from a bad GPS signal.
- It’s important to check that the measurements roughly follow a Gaussian distribution.