SLT Low 1: Introduction to Statistical Learning and Its Geometry

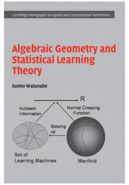
Edmund Lau

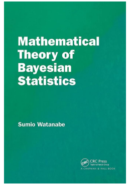
Overview of SLT Low Road

- **SLT Low 1:** Introduction to statistical learning theory and its geometry.
- **SLT Low 2:** Bridging regular and singular models.
- **SLT Low 3:** Introduction to Algebraic Geometry, blowups, resolution of singularities and computing RLCTs.
- **SLT Low 4:** Proof sketch of the free energy formula.

Main references:

Gray book and **Green book** by Sumio Watanabe.


Cambridge Monographs on Applied and Computational Mathematics
Algebraic Geometry and Statistical Learning Theory
Sumio Watanabe
R
Kullback Information
Normal Crossing Function
Blowing up
Set of Learning Machines
Manifold



Goals

- Introduce background materials for Singular Learning Theory (henceforth **SLT**).
- Standardise notations / conventions / notion of “what learning is”.
- Illustrate effects of geometry in learning machines.

This session serves as an introduction to Statistical Learning.

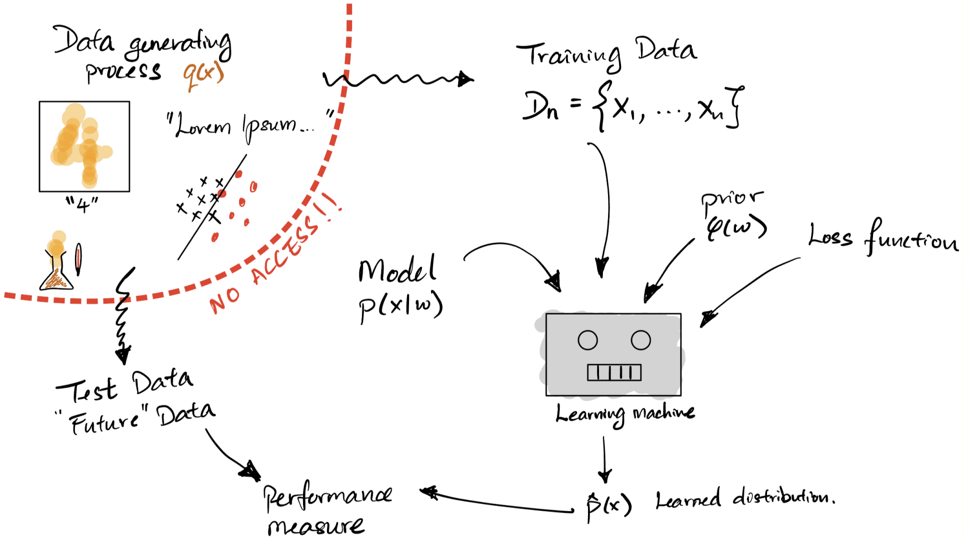
There are a lot of background definitnions, concepts and notations required for SLT. Listing them will be daunting, but we will go over them by working with them in simple examples.

We will also illustrate various important phenomena in statistical learning, particulary the effect of geometry.

Furthermore, we want to

1. Bring relevant knowledge back to your mental cache.
2. Use some notation often enough that they get store in your mental cache.
3. Might have bits and pieces that are new to you.
4. Won’t be so detailed that we discuss at the level of measure theoretic issues though.

Cartoon Picture: Statistical Learning



High-level picture of what a statistical learning problem looks like. These notions are more or less accepted in the literature with some variation.

1. Introduce what we mean by a "Learning Machine", an effectively computable measureable function sending model and data to an output distribution that well-approximates the true distribution.
2. Discuss various components of a statistical learning problem.
3. Discuss variations and examples.

Notations

Just absorb by osmosis for a moment ...

True distribution $q(x)$, $x \in \mathbb{R}^N$.

Training Data $D_n = \{X_1, X_2, \dots, X_n\}$, $X_i \sim q(x)$ i.i.d.

Model $p(x \mid w)$, where $w \in W \subset \mathbb{R}^d$ with prior $\varphi(w)$ on W.

Likelihood $p(D_n \mid w) = \prod_{i=1}^{n} p(X_i \mid w)$.

Negative log-likelihood $L_n(w) = \frac{1}{n} \sum_{i=1}^{n} \log p(X_i \mid w)$

Averaged NLL $L(w) = \int q(x) \log p(x \mid w) dx$

$\mathrm{LearningMachine} : (p(x \mid w), D_n) \mapsto \hat{p}(x)$.

Choose *Kullback-Leibler divergence* as "closeness" measure between distribution

$\mathrm{D}_{KL}\left(q(x) \| \hat{p}(x)\right) := \int_X q(x) \log \frac{q(x)}{\hat{p}(x)} dx$.

This is mainly to flash all important notation at once and then illustrate them with examples later.

Example: Coin Toss

True distribution: $q(x) = \begin{cases} w_0 & \text{if } x = \mathrm{H} \\ 1 - w_0 & \text{if } x = \mathrm{T} \end{cases}, \; w_0 \in [0, 1].$

Data: $D_n = \{\mathrm{H}, \mathrm{T}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{T}, \mathrm{T}, \dots, \mathrm{T}\}.$

Simple model: $p(x \mid w) = \begin{cases} w & \text{if } x = \mathrm{H} \\ 1 - w & \text{if } x = \mathrm{T} \end{cases}, \; w \in [0, 1].$

Negative log-likelihood:

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i \mid w) = \frac{h}{n}\log w + \frac{t}{n}\log(1 - w)$$

Which converge by Law of Large number to its mean:

$$L(w) = -\sum_{x=\mathrm{H},\mathrm{T}} q(x)\log p(x \mid w) = -w_0 \log(w) - (1 - w_0)\log(1 - w)$$

Under the assumption that seeing definintion and concepts in action serves as better introduction, we shall go through this "Statistics 101" type problem using MLE, where everything goes *right*.

It serves as a pretext to cram a bunch of concept in statistical learning in one go... sorry... $p(D_n \mid w)$, $L_n(w)$, $L(w)$, $\mathrm{I}(w)$, MLE, unbiased estimator, asymptotic consistency / efficiency / normality, Fisher Information ...

We could ask for $p(x \mid \mathrm{argmin} L(w))$ which will simply recover $\hat{p}(x) = q(x)$ for us. But that's not learning!

Example: Coin Toss

Maximum Likelihood Estimator $\hat{w}_{\mathrm{MLE}} = \mathrm{argmin}_w L_n(w)$:

$$\partial_w L_n(\hat{w}) = 0 \implies \frac{h}{\hat{w}} - \frac{t}{1-\hat{w}} = 0 \implies \hat{w} = \frac{h}{n} = 1 - \frac{t}{n}$$

We now have the learning machine:

$$\mathrm{CoinTossMLEMachine} : D_n \mapsto \hat{p}_{MLE}(x) = p\left(x \mid \frac{h}{n}\right).$$

Introduce a conceptually simple learning method, MLE:

- $\hat{w}$ is a random variable depending on D_n.
- Already there are some geometric issue at the boundary when $\hat{w} = 0$ or $h, t = 0$.
- Checking second order condition shows that this is indeed a local (indeed unique global) minimum.

With the learned distribution, we can answer things like:

- Given observed data, what is the chance of the next toss landing head?
- Given historical record, how much should one bet on H to maximise return in the long run?

Example: Coin Toss

How well does it do?

A: We can compute the distribution of $\hat{w}_{\mathrm{MLE}}$ *exactly* in this case (!)

$$p\left(\hat{w} = \frac{h}{n} \;\middle|\; w_0\right) = \binom{n}{h} w_0^h (1 - w_0)^{n-h}$$

With this we can ask:

- What is the expected prediction? A: $\mathbb{E}\left[\hat{w}\right] = w_0$
- What is the variance of the prediction? A: $\mathbb{V}\left(\hat{w}\right) = \frac{w_0(1-w_0)}{n}$
- What is the asymptotic behaviour as $n \to \infty$? A: $\hat{w} \sim N(w_0, \frac{w_0(1-w_0)}{n})$

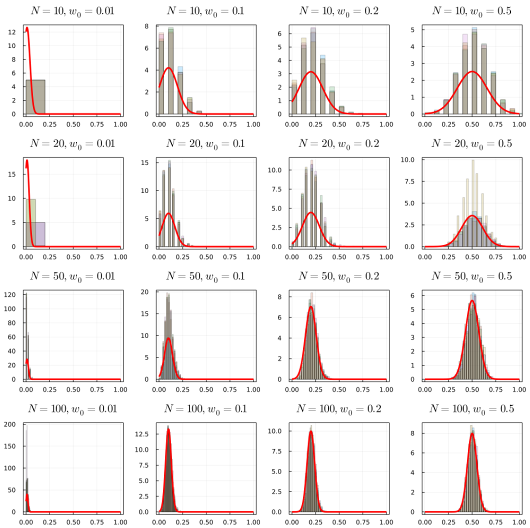
Introduce various ways of answering "how well does it do?"

- How different is $\hat{p}(x)$ from $q(x)$?
- What is the expected prediction? (unbiased in this case)
- How does this prediction change with D_n?
- If I collect more data, does my prediction improve? By how much?

Introduce Kullback-Leibler Divergence.

Asymptotic Normality.

Coin toss MLE distributions



Explain the plot. x-axis = MLE $\simeq$ Output of CoinTossMLEMachine = learned distribution. Repeat experiment / learning problem lots of time, with different D_n drawn, to get a distribution.

Top to bottom: Increasing n. Left to right: different true parameter.

Fisher Information Matrix

How difficult is it to narrow down to the true w_0?

That depends on where w_0 is and the geometry near w_0!

Definition (Fisher Information Matrix)

Let $p(x \mid w)$ be a model, the Fisher information matrix $\mathrm{I}(w)$ at $w \in W$ is a matrix with the ij^{th} components

$$[\mathrm{I}(w)]_{ij} := \int_X \left(\partial_{w_i} \log p(x \mid w)\right) \left(\partial_{w_j} \log p(x \mid w)\right) p(x \mid w) dx.$$

For the coin toss model:

$$\begin{aligned} \mathrm{I}(w) &= \sum_{X=\mathrm{H},\mathrm{T}} \left(\frac{d}{dw} \log p(x \mid w)\right)^2 p(x \mid w) \\ &= \frac{1}{w^2} w - \frac{1}{1-w}(1-w) = \frac{1}{w(1-w)} \end{aligned}$$

Introduce Fisher informatin and Asymptotic efficiency.

Curvature near w_0 determine how many samples is needed to eliminate hypothesis "close to" w_0. Already, the geometry (differential geometry here) of the model is influencing learning behaviour!!

Example: Coin Toss

Theorem (Cramer-Rao inequality)

Let $\hat{w}$ is an unbiased estimator for a model $p(x \mid w)$ computed from a set of i.i.d. samples $\{X_1, \ldots X_n\}$, $X_i \sim p(x \mid w)$. Let $V_{ij} = \mathbb{E}\left[(\hat{w}_i - w_i)(\hat{w}_j - w_j)\right]$ be the covariance matrix and $\mathrm{I}(w)$ the Fisher information matrix for the model. Then $V \geqslant (n\mathrm{I}(w))^{-1}$.

For $\mathrm{CoinTossMLEMachine}$, we have

$$V(\hat{w}_{MLE}) = \frac{1}{n\mathrm{I}(w_0)}$$

$\implies$ $\mathrm{CoinTossMLEMachine}$ is *efficient*.

In this case, MLE is best among unbiased estimator. You can't narrow down near w_0 faster than $\hat{w}_{MLE}$ if you want to maintain $\mathbb{E}[\hat{w}] = w_0$. This is known as efficiency. Regular models are asymptotically efficient.

It doesn't mean you couldn't do better if you allow for biased estimators.

Inequality of matrices is defined as $A \leqslant B \iff A - B$ is positive semidefinite.

Example: Simple 2D Gaussian

True distribution: $q(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-a_0)^2+(y-b_0)^2}{2}\right)$.

Data: $D_n = \{(X_i, Y_i)\}_i = \{(0.12, 0.1), (-0.2, 3.14) \dots\}$.

Model: $p(x, y \mid w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-a)^2+(y-b)^2}{2}\right)$, $w = (a, b)$.

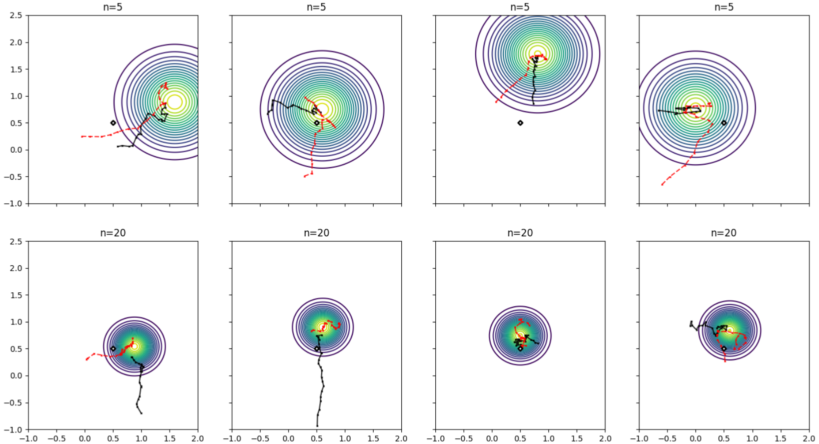
Negative log-likelihood:

$$L_n(w) = \frac{1}{2n}\sum_{i=1}^{n}\left[(X_i - a)^2 + (Y_i - b)^2\right] - \frac{1}{2}\log(2\pi)$$

Another regular model for variety. Also illustrating that a model can be used with different learning method.

There is a L^2 loss term, or like a simple harmonic oscilator.

Example: Gaussian $N(\mu, \sigma^2)$



Likelihood landscape.

Illustrates other learning machine: MLE, SGD, SGD + momentum, different number of epoch, etc

Introduce posterior and Bayes predictive distribution and Gibbs learning.

Summarising

$q(x)$, D_n, X_i, $w \in W$, $p(x \mid w)$, $L_n(w)$, $L(w)$, D_{KL}, $\mathrm{I}(w), \ldots$

- Models can be used with different "training algorithm".
- Predictions are random variables that depend on data D_n.
- Performance measures of the learned quantities are random variables too.
- That is why we need to study their *statistics* (expectation, variance, etc).

In the preceeding models, things are as nice as possible.

- Identifiable.
- Positive definite Fisher information matrix, $\mathrm{I}(w)$.
- $L(w)$ has a unique global minimum.
- Asymptotically consistent, normal and efficient.

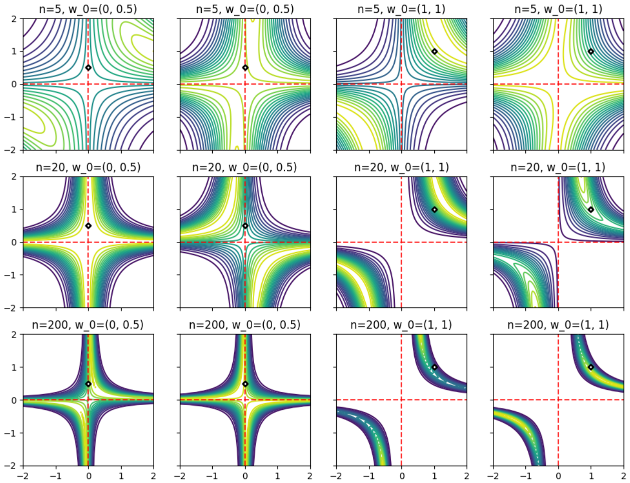
We get a taste of what we care about in statistical learning.

What kind of questions are raised and what flavor are their answers.

Many of these *regularity* assumptions are implicit throughout statistics and sometimes they are uncritically assumed for models where they do not apply, like most ML models.

But ...

Example: $y = a\tanh(bx) + \eta$



Here's a likelihood contour of a model.

Obervations:

- Different w_0 seems to completely change the shape of W_0.
- Even at high n, it is no where close to being Gaussian.
- The changes in $L_n(w)$ as we resample D_n is not just a change in location, it is more dramatic than that.

What's going on ...?

Example: Simple 2D Gaussian

Regression model: $y = a\tanh(bx) + \eta$, where $\eta \sim N(0,1)$

$$p(y \mid x, w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(y - a\tanh(bx)\right)^2\right).$$

If the true regression function is $y = f_{true}(x) = a_0 \tanh(b_0 x)$

Fix input distribution $q(x)$,

$$D_n = \{(X_i, Y_i) \mid X_i \sim q(x),\ Y_i = f_{true}(X_i)\}$$

$$L_n(w) = \frac{1}{2n}\sum_i \left(Y_i - a\tanh(bX_i)\right)^2 + \text{const}$$

$$L(w) = \frac{1}{2}\int q(x)\left(a\tanh(bx) - a_0\tanh(b_0x)\right)^2 dx + \text{const}$$

This is an example of a model with *singularity*. What do we mean by that?

Observe that the minima of $L(w)$ is no longer a single point. And depending on what w_0 is, it is not even a smooth manifold!

When we say $\mathrm{I}(w)$ controls the curvature and how distinguishable distributions in a neighbourhood of w is? Well, that characterisation requires that $\mathrm{I}(w)$ be *non-degenerate*. And in this case, it is degenerate near $ab = 0$.

Regular vs Singular Models

Definition (Singular Models)

A statistical model $p(x \mid w)$ is **regular** if

- It is **identifiable**. This means that the map $W \ni w \mapsto p(- \mid w)$ is injective. In other words, if $p(x \mid w_1) = p(x \mid w_2)$ almost everywhere, then $w_1 = w_2$.
- It has **everywhere positive definite Fisher information matrix**. Since $\mathrm{I}(w)$ is always positive semi-definite, this is equivalent to having invertible $\mathrm{I}(w)$.

A statistical model is **strictly singular** if it is not regular.

Singular models encompass both regular and strictly singular models.

Give definitions of regular and singular models. Discuss relative position of true parameter and the singular set.

Pause

Any questions so far...?

Bayesian Learning

Central Objects in Bayesian Statistics and SLT

Given a truth-model-prior triplet, $(q(x), p(x \mid w), \varphi(w))$

And i.i.d. dataset $D_n = \{X_i\}$, $X_i \sim q$,

Negative log-likelihood: $L_n(w) = -\frac{1}{n}\sum_i \log p(x \mid w)$

Boltzmann weight: $e^{-nL_n(w)}\varphi(w)$

Partition function / Evidence / Marginal Likelihood:

$Z_n = \int_W e^{-nL_n(w)}\varphi(w)dw$

Posterior distribution:

$$\begin{aligned} p(w \mid D_n) &= \frac{p(D_n \mid w)\varphi(w)}{p(D_n)} && \text{Bayes theorem} \\ &= \frac{1}{Z_n} e^{-nL_n(w)}\varphi(w) \end{aligned}$$

Our goal is to understand singularities of learning machines. The machinery of SLT allow us to do that. But explicit theorems of SLT focus on Bayesian Learning.

Although we focus on Bayesian Learning, SLT allows us to study the singularity of the model which might be used in different learning method, e.g. MLE, MAP, SGD.

Example: regression $y = h_w(x) +$ noise

Suppose $\varphi(w) \propto e^{-\gamma \|w\|^2}$

Gaussian noise:

$$p(y \mid x, w) \propto e^{-\frac{1}{2\sigma^2}(y - h_w(x))^2}$$

$$L_n(w) = -\frac{1}{2\sigma^2 n} \sum_i (Y_i - h_w(X_i))^2$$

$$p(w \mid D_n) = \frac{1}{Z_n} e^{-\frac{1}{2\sigma^2}\mathrm{MSE}(w) - \gamma \|w\|^2}$$

Exponetially distributed noise:

$$p(y \mid x, w) \propto e^{-\frac{1}{b}\mid y - h_w(x) \mid}$$

$$L_n(w) = -\frac{1}{bn} \sum_i \mid Y_i - h_w(X_i) \mid$$

$$p(w \mid D_n) = \frac{1}{Z_n} e^{-\frac{1}{b}\mathrm{MAE}(w) - \gamma \|w\|^2}$$

Normalised Quantities

For theoretical discussion of learning machines behaviour we introduce quantities that are **normalised**.

$$\begin{aligned}
L_n(w) &\quad\rightsquigarrow\quad K_n(w) = L_n(w) - S_n \\
e^{-nL_n(w)}\varphi(w) &\quad\rightsquigarrow\quad e^{-nK_n(w)}\varphi(w) \\
Z_n &\quad\rightsquigarrow\quad \overline{Z}_n = Z_n / e^{-nS_n} = \int_W e^{-nK_n(w)}\varphi(w)dw \\
F_n &\quad\rightsquigarrow\quad \overline{F}_n = -\log \overline{Z}_n
\end{aligned}$$

The posterior remains

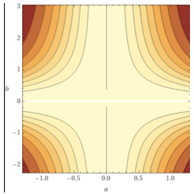
$$p(w \mid D_n) = \frac{1}{Z_n} e^{-nL_n(w)}\varphi(w) = \frac{\cancel{e^{-nS_n}}}{\cancel{e^{-nS_n}}} \frac{1}{\overline{Z}_n} e^{-nK_n(w)}\varphi(w)$$

The log-likelihood acts like a “Energy” function in a disordered system. The posterior distribution is analogous to the thermal equilibrium distribution of the state variables w.

Separating Effects of Geometry and Stochasticity

$$
\begin{aligned}
L_n(w) &\rightsquigarrow L(w) = \mathbb{E}\left[L_n(w)\right] = -\int q(x) \log p(x \mid w) dx \\
e^{-nL_n(w)}\varphi(w) &\rightsquigarrow e^{-nL(w)}\varphi(w) \\
Z_n &\rightsquigarrow Z(n) = \int_W e^{-nL(w)}\varphi(w) dw \\
F_n &\rightsquigarrow \overline{F}_n = -\log Z(n)
\end{aligned}
$$

Determistic likelihood for the 2D-tanh model with $w_0 = (0, 0)$.



“Anealed” quantities.

Arrangement of concepts

$$\begin{tikzcd}
\boxed{\begin{matrix} L_n(w),\ e^{-nL_n(w)}, \\ Z_n,\ F_n \end{matrix}} \arrow[r, "\text{normalise}"] \arrow[d, "\text{``anneal''}"'] & \boxed{\begin{matrix} K_n(w), \\ e^{-nK_n(w)},\ \overline{Z}_n,\ \overline{F}_n \end{matrix}} \arrow[d, "\text{``anneal''}"] \\
\boxed{\begin{matrix} L(w),\ e^{-nL(w)}, \\ Z(n),\ F(n) \end{matrix}} \arrow[r, "\text{normalise}"] & \boxed{\begin{matrix} K(w),\ e^{-nK(w)}, \\ \overline{Z}(n),\ \overline{F}(n) \end{matrix}}
\end{tikzcd}$$

This is how I (informally) arrange these quantities in my mind. Might be helpful to use the right quantity in the right context (theoretical, experimental, geometric, stochastic...)

Brief Digression: Realisability

We have been implicitly assuming that $q(x) = p(x \mid w_0)$ for some w_0. This is the **realisability** assumption.

In practice, the data might not come from a distribution contained in $\{p(x \mid w)\}$, or any finite dimensional distribution at all.

In theoretical analysis, instead of normalising by log-likelihood at a true parameter, $w_0 \in W_0$, we use an *optimal parameter* $w_0 = W_{opt} = \{w_0 : L(w_0) = \min_w L(w)\}$

$$K_n(w) = L_n(w) - L_n(w_0), \quad K(w) = L(w) - L(w_0).$$

With some extra assumption, the theory then proceed in similar fashion.

Discuss "realisability" condition. What changes in the unrealisable case? Discuss hypothesis testing / model selection where we don't we don't care much about the true distribution. In these cases, it is clear that W_{crit} is important even if it has measure zero.

Indeed, we are interested in all critical level sets of $L(w)$. We shall discuss this when we talk about **phase transition**.

Bayes Learning Machine

Bayesian predictive distribution:

$$\mathrm{BayesMachine} : (p(x \mid w), D_n) \mapsto \hat{p}_{Bayes}(x)$$
$$\hat{p}_{bayes}(x) := \int_W p(x \mid w) p(w \mid D_n) dw$$

Errors (theoretical performance measures):

$$B_g(n) = \mathrm{D}_{KL}(q(x) \| \hat{p}_{Bayes}(x)) = \int q(x) \log \frac{q(x)}{p(x \mid D_n)} dx$$
$$B_t(n) = \frac{1}{n} \sum_i \frac{q(X_i)}{p(X_i \mid D_n)}$$

“Bayes Quartet”

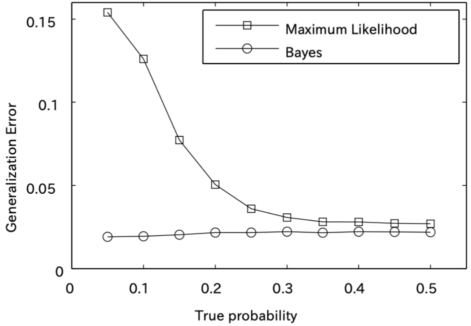
Preview

State FEF.

State how we will get there: Geometry first and then consider stochastic effects.

Bayesian Learning on Coin Toss Problem

Let's work through a $\mathrm{CoinTossBayesMachine}$.



This is Exercise 1.9.3 in the green book.

Illustrate the generalisation error of Bayesian vs MLE.

Next time: A bunch of integrals.