# Lecture 2b. Statistical Schools of Thought: Statistical Decision Theory

COMP90051 Statistical Machine Learning

Semester 2, 2020
Lecturer: Ben Rubinstein

THE UNIVERSITY OF MELBOURNE

# This lecture

## How do learning algorithms come about?

- Frequentist statistics
- **Statistical decision theory**
- **Extremum estimators**
- Bayesian statistics

## Types of probabilistic models

- Parametric vs. Non-parametric
- Generative vs. Discriminative

# Statistical Decision Theory

Branch within statistics, optimisation, economics, control, emphasising utility maximisation.

# Decision theory

Wald

- Act to maximise utility - connected to economics and operations research
- Decision rule $\delta(\boldsymbol{x}) \in A$ an action space
  * E.g. Point estimate $\hat{\theta}(x_1, \ldots, x_n)$
  * E.g. Out-of-sample prediction $\hat{Y}_{n+1}|X_1, Y_1, \ldots, X_n, Y_n, X_{n+1}$
- Loss function $l(a, \theta)$: economic cost, error metric
  * E.g. square loss of estimate $\left(\hat{\theta} - \theta\right)^2$
  * E.g. 0-1 loss of classifier predictions $1[y \neq \hat{y}]$

# Risk & Empirical Risk Minimisation (ERM)

- In decision theory, really care about *expected* loss
- Risk $R_\theta[\delta] = \mathrm{E}_{\boldsymbol{X}\sim\theta}[l(\delta(\boldsymbol{X}),\theta)]$
  - E.g. true test error
  - aka generalization error
- Want: Choose $\delta$ to minimise $R_\theta[\delta]$
- Can't directly! Why?
- ERM: Use training set ***X*** to approximate $p_\theta$
  - Minimise empirical risk $\hat{R}_\theta[\delta] = \frac{1}{n}\sum_{i=1}^{n} l(\delta(X_i),\theta)$

# Decision theory vs. Bias-variance

We've already seen

- Bias: $\mathrm{B}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[\hat{\theta}(X_1, \dots, X_n)\right] - \theta$
- Variance: $\mathrm{Var}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[(\hat{\theta} - \mathrm{E}_\theta[\hat{\theta}])^2\right]$

But are they equally important? How related?

- Bias-variance decomposition of square-loss risk

$$\mathrm{E}_\theta\left[\left(\theta - \hat{\theta}\right)^2\right] = [\mathrm{B}(\hat{\theta})]^2 + \mathrm{Var}_\theta(\hat{\theta})$$

# Extremum estimators

Very general framework that covers elements of major statistical learning frameworks; enjoys good asymptotic behaviour in general*!!*

# Extremum estimators

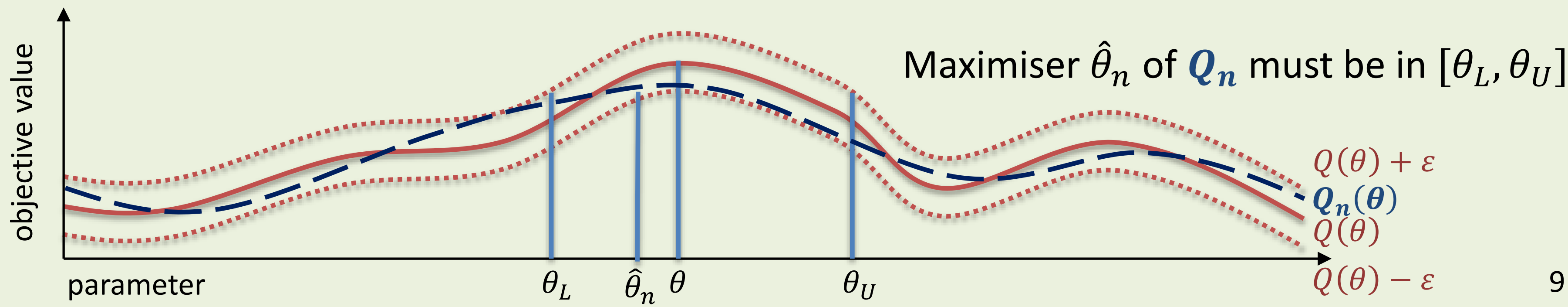
- $\hat{\theta}_n(\boldsymbol{X}) \in \underset{\theta\in\Theta}{\operatorname{argmin}}\, Q_n(\boldsymbol{X}, \theta)$ for any objective $Q_n()$

- Generalises bits of all statistical frameworks. *W00t!*
  - MLE and ERM seen earlier this lecture; and
  - MAP seen later in this lecture.
  - These are all *M*-estimators, with *Q* as a sum over data (i.e. of log-likelihood, loss, or log-likelihood plus log prior)

- And it generalises other frameworks too!

# Consistency of Extremum Estimators


- Recall consistency: stochastic convergence to 0 bias
- Theorem for extremum estimators: $\hat{\theta}_n \to \theta$ in prob, if there's a ("limiting") function *Q*() such that:
  1. *Q*() is uniquely maximised by $\theta$.
     That is, no other parameters make *Q*() as large as $Q(\theta)$.
  2. The parameter family $\Theta$ is "compact"
     (a generalisation of the familiar "closed" & "bounded" set, like [0,1])
  3. *Q*() is a continuous function
  4. Uniform convergence: $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \to 0$ in probability.



# A game changer

- Frequentists: estimators that aren't even correct with infinite data (inconsistent), aren't adequate in practice
- Proving consistency for every new estimator? Ouch!

- So many estimators are extremum estimators – general guarantees make it much easy (but not easy!) to prove
- Asymptotic normality
  - Extremum estimators converge to Gaussian in distribution
  - Asymptotic efficiency: the variance of that limiting Gaussian
- Practical: Confidence intervals - think error bars*!!*

→ ***Frequentists like to have this asymptotic theory for their algorithms***

# Summary

- Decision theory: Utility-based, Minimise risk
- Many familiar learners minimise loss over data (ERM)
- Extremum estimators generalise ERM, MLE, (later: MAP)
  * Amazingly, consistent: Gives us confidence that they work (eventually)
  * Amazingly, asymptotically normal: Helps make confidence intervals

Next time: Last but not least, the Bayesian paradigm

Workshops week #2: learning Bayes one coin flip at a time!