Statistical Learning

Lecture 1:
Statistical Decision and Estimation Theory

Mário A. T. Figueiredo

Instituto Superior Técnico & Instituto de Telecomunicações

University of Lisbon, **Portugal**

March 2024

Statistical Decision Theory

- Specify the set of possible *decisions/action*

  Examples: classify email as spam or not; predict the next word in a text.

- Specify possible underlying truth (*state of nature*)

  Examples: email message is/isn't spam; actual word.

- Specify *observations*.

  Example: message and metadata (`from`, `subject`, `IP`, `...`); context text.

- Specify how *observations* are related to *state of nature*.

  Example: word statistics.

- Specify payoff/consequences of the decision.

  Example: *loss* of sending a good email to the spam folder (vice-versa).

- **Goal**: derive optimal decision rules; characterize decision rules.

Statistical Decision Theory

- Unknown underlying reality/truth, the *state of nature*: $s \in \mathcal{S}$
- *Observations*: a random variable $X \in \mathcal{X}$, with known $f_X(\cdot|s)$;
  It can be a pdf, pmf, hybrid, univariate, multivariate,...
  Often called likelihood function
- Set of allowed *decisions/action* $\mathcal{A}$
- *Decision rules* $\delta : \mathcal{X} \to \mathcal{A}$, where $\delta \in \mathcal{D}$:

$$\delta(x) = a \quad \text{means "if } x \text{ is observed, decide } a\text{"}$$

- *Loss function*, quantifying the decision consequences, $L : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$L(s, a) = \text{"the loss of deciding } a \text{ under state of nature (truth) } s\text{"}$$

Statistical decision problem: given $(\mathcal{S}, \mathcal{A}, \mathcal{X}, f_X, \mathcal{D}, L)$,
choose optimal $\delta \in \mathcal{D}$.

Statistical Decision Problems: Examples

- Medical decision problem: $\mathcal{S} = \{\text{cancer}, \text{no cancer}\}$,
  $\mathcal{A} = \{\text{surgery}, \text{chemotherapy}, \text{do nothing}\}$,
  $X = $ lab test results, images, ...
- Fingerprint id system: $\mathcal{S} = \{\text{user 1}, \text{user 2}, ..., \text{user N}, \text{stranger}\}$
  $\mathcal{A} = \{\text{login user 1}, ..., \text{login user N}, \text{not allowed}\}$
  $X = $ data aquired at the fingeprint sensor
- Parameter estimation: $\mathcal{S} = \mathcal{A} = \mathbb{R}$
  $X = $ observations; known $f_X(\cdot|s)$, for $s \in \mathcal{S}$
- Binary decision: $\mathcal{S} = \mathcal{A} = \{0, 1\}$
  $X = $ observed data; $f_X(\cdot|s)$ observation model.
- Binary decision with reject option: $\mathcal{S} = \{0, 1\}$; $\mathcal{A} = \{0, 1, R\}$

Critical aspect: (unrealistically) assumes perfect knowledge of f_X

Statistical Decision Problems: Frequentist Risk

Assumptions: $s \in \mathcal{S}$ is fixed but unknown
decisions should have low expected (w.r.t. $X \sim f_X$) loss

The *risk:* $R(s, \delta) = \mathbb{E}_X \left[L(s, \delta(X)) | s \right]$

$$\mathbb{E}_X \left[L(s, \delta(X)) | s \right] = \begin{cases} \displaystyle\int_{\mathcal{X}} L(s, \delta(x)) \, f_X(x|s) \, dx & \Leftarrow \quad \mathcal{X} \text{ is continuous} \\ \displaystyle\sum_{x \in \mathcal{X}} L(s, \delta(x)) \, f_X(x|s) & \Leftarrow \quad \mathcal{X} \text{ is discrete;} \end{cases}$$

Question: how to use the risk to compare/order different decisions?
The risk depends on s, and s is unknown!

Domination and Admissibility

The *risk:* $R(s, \delta) = \mathbb{E}_X \left[L(s, \delta(X)) | s \right]$

...induces a partial order: domination

Definition: A rule $\delta \in \mathcal{D}$ is dominated by another rule $\delta' \in \mathcal{D}$ if

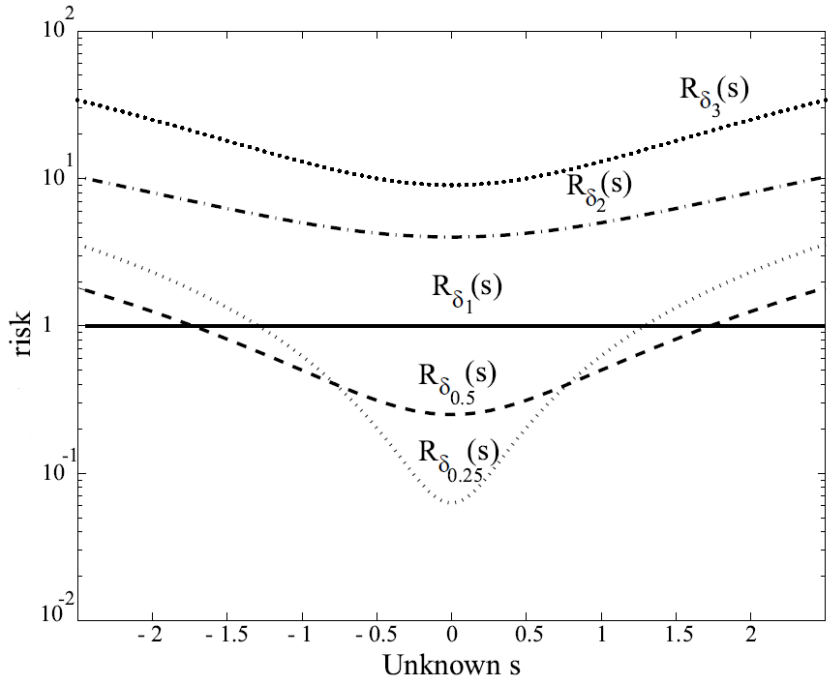
(i) for any $s \in \mathcal{S}$, $R(s, \delta') \leq R(s, \delta)$, and

(ii) there exists $s_0 \in \mathcal{S}$ such that $R(s_0, \delta') < R(s_0, \delta)$.

Definition: a rule $\delta \in \mathcal{D}$ is admissible if it is not dominated by any other rule in $\mathcal{D}$

Admissible Rules: Parameter Estimation Example

- unknown real parameter $s \in \mathcal{S} = \mathbb{R}$, to be estimated ($\mathcal{A} = \mathcal{S} = \mathbb{R}$)
- single Gaussian observation $X \in \mathbb{R}$, with $f_X(x) = \mathcal{N}(x; s, 1)$
- allowed decision rules are linear: $\mathcal{D} = \{\delta_k(x) = k\,x,\ k \in \mathbb{R}_+\}$
- loss function: $L(s, a) = (s - a)^2$
- the risk is $R(s, \delta_k) = \mathbb{E}\big[(s - k\,X)^2 | s\big] = s^2(1 - k)^2 + k^2$



Any δ_k, for $k > 1$, is not admissible. All δ_k, for $k \leq 1$, are not ordered

Minimax Rules

Definition: minimax risk:

$$\min_{\delta \in \mathcal{D}} \max_{s \in \mathcal{S}} R(s, \delta)$$

(with $\inf$ and $\sup$ in case $\mathcal{D}$ and $\mathcal{S}$ are open)

Definition: δ^* is a minimax rule if

$$\max_{s \in \mathcal{S}} R(s, \delta^*) = \min_{\delta \in \mathcal{D}} \max_{s \in \mathcal{S}} R(s, \delta)$$

Back to the previous estimation example: $R(s, \delta_k) = s^2(1-k)^2 + k^2$

$$\sup_{s \in \mathbb{R}} R(s, \delta_k) = \sup_{s \in \mathbb{R}} \left\{ s^2(1-k)^2 + k^2 \right\} = \begin{cases} 1 & \Leftarrow \quad k = 1 \\ \infty & \Leftarrow \quad k \neq 1 \end{cases}$$

...thus δ_1 is the (unique) minimax rule.

Frequentist Characterization of Estimators

Estimation problems ($\mathcal{S} = \mathcal{A} = \mathbb{R}$): goal is to estimate $s \in \mathbb{R}$.

Observations: $X = (X_1, ..., X_n) \in \mathbb{R}^n$, with $f_X(\cdot|s)$

- Consistent estimator: $\lim_{n\to\infty} \delta(X) = s$, in probability ($\delta(X)$ is a r.v.), *i.e.*,

  $\forall_{\varepsilon>0}, \lim_{n\to\infty} \mathbb{P}\big(|\delta(X) - s| > \varepsilon\big) = 0$

- Bias of an estimator: $\mathbb{E}\big[\delta(X)\big] - s$ (if $\mathbb{E}\big[\delta(X)\big] = s$, δ is unbiased)

- Variance of an estimator: $\mathbb{E}\Big[\Big(\delta(X) - \mathbb{E}[\delta(X)]\Big)^2\Big]$

- MVU *(minimum variance unbiased)* estimator, $\delta_{\text{MVU}} \in \mathcal{D}$

  $$\mathbb{E}[\delta_{\text{MVU}}(X)] = s \quad \text{and} \quad \mathbb{E}\Big[\big(\delta_{\text{MVU}}(X) - s\big)^2\Big] \leq \mathbb{E}\Big[\big(\delta(X) - s\big)^2\Big]$$

  for any other unbiased $\delta \in \mathcal{D}$. Also known as efficient estimator.

Frequentist Characterization of Estimators: Example 1

Independent and identically distributed (i.i.d.) observations: $X = (X_1, ..., X_n)$,

$$f_X(x|s) = \prod_{i=1}^{n} \mathcal{N}(x_i; s, \sigma^2) = \left(\frac{1}{2\sigma^2}\right)^2 \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - s)^2\right)$$

Sample mean estimator: $\delta(x) = \dfrac{x_1 + ... + x_n}{n}$

- Unbiased: $\mathbb{E}[\delta(X)] = \dfrac{1}{n}\left(\mathbb{E}(X_1) + ... + \mathbb{E}(X_n)\right) = s$
- Consistent: $\delta(X) \sim \mathcal{N}(s, \phi_n^2)$,

$$\phi_n^2 = \text{var}\left(\frac{X_1}{n} + ... + \frac{X_1}{n}\right) = n\,\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$$

- Minumum variance? ...more later.

Frequentist Characterization of Estimators: Example 2

Observations: $X = (X_1, ..., X_n) \in \mathbb{R}^n$, where $s = (s_1, s_2) = (\mu, \sigma^2)$ and

$$f_X(x|s) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2) = \left(\frac{1}{2\sigma^2}\right)^2 \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

Sample variance: $\delta_2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \delta_1(x))^2$, where $\delta_1(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$

We have seen that δ_1 is unbiased and consistent; what about δ_2?

- $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j)^2\right] = \frac{n-1}{n}\sigma^2$ (using $\mathbb{E}[X_i^2] = \mu^2 + \sigma^2$; $\mathbb{E}[X_i X_j] = \mu^2$)
- Unbiased variance estimator: $\hat{\sigma}^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2$

Minimum Variance and the Cramér-Rao Bound

Let $s \in \mathbb{R}$ be a (scalar) parameter to be estimated from $X \sim f_X(\cdot|s)$

Cramér-Rao bound:

$$\underbrace{\mathbb{E}[\delta(X)] = s}_{\delta \text{ is unbiased}} \Rightarrow \text{var}[\delta(X)] \equiv \mathbb{E}[(\delta(X) - s)^2] \geq \frac{1}{I(s)},$$

Fisher information: $I(s) = -\mathbb{E}\left[\frac{\partial^2 \log f_X(X|s)}{\partial s^2}\right]$ $\quad (\log = \log_e)$

(curvature)

Example: $X = (X_1, ..., X_n) \in \mathbb{R}^n$, with $f_X(x|s) = \prod_{i=1}^{n} \mathcal{N}(x_i; s, \sigma^2)$

$$-\log f_X(x|s) = K + \frac{\sum_{i=1}^{n}(X_i - s)^2}{2\sigma^2} \Rightarrow I(s) = \frac{\mathbb{E}\left[\frac{\partial^2 \sum_{i=1}^{n}(s-X_i)^2}{\partial s^2}\right]}{2\sigma^2} = \frac{n}{\sigma^2}$$

thus, since $\delta(x) = \frac{x_1 + ... + x_n}{n}$ is unbiased and has variance $\frac{\sigma^2}{n}$, it is MVU

Cramér-Rao Bound: Proof (simplified)

- Score: $V(X) = \dfrac{\partial \log f_X(X|s)}{\partial s} = \dfrac{1}{f_X(X|s)} \dfrac{\partial f_X(X|s)}{\partial s}$ (random)
- $\mathbb{E}[V(X)] = \displaystyle\int_{\mathcal{X}} f_X(x|s) \frac{1}{f_X(x|s)} \frac{\partial f_X(x|s)}{\partial s} dx = \frac{\partial}{\partial s} \int_{\mathcal{X}} f_X(x|s) dx = 0$
- $\text{cov}(V(X), \delta(X)) = \mathbb{E}[V(X)\delta(X)] - \underbrace{\mathbb{E}[V(X)]}_{0} \mathbb{E}[\delta(X)] = \mathbb{E}[V(X)\delta(X)]$
- $\mathbb{E}[V(X)\delta(X)] = \displaystyle\int_{\mathcal{X}} \delta(x) \frac{\partial f_X(x|s)}{\partial s} dx = \frac{\partial}{\partial s} \underbrace{\int_{\mathcal{X}} \delta(x) f_X(x|s) dx}_{\mathbb{E}[\delta(X)]=s} = 1$
- $\text{cov}(V(X), \delta(X)) = 1$

Cramér-Rao Bound: Proof (continuation)

- Fisher information is the negative score variance:

$$
\begin{aligned}
I(s) &= -\mathbb{E}\left[\frac{\partial^2 \log f_X(X|s)}{\partial s^2}\right] \\
&= -\mathbb{E}\left[\frac{\partial}{\partial s}\Big(\frac{1}{f_X(X|s)}\frac{\partial f_X(X|s)}{\partial s}\Big)\right] \\
&= \mathbb{E}\left[\frac{1}{\big(f_X(X|s)\big)^2}\Big(\frac{\partial f_X(X|s)}{\partial s}\Big)^2\right] - \mathbb{E}\left[\frac{1}{f_X(X|s)}\frac{\partial^2 f_X(X|s)}{\partial s^2}\right] \\
&= \underbrace{\mathbb{E}\left[\Big(\frac{\partial \log f_X(X|s)}{\partial s}\Big)^2\right]}_{\mathsf{var}(V(X))} - \underbrace{\frac{\partial^2}{\partial s^2}\int_{\mathcal{X}} f_X(x|s)\,dx}_{0}
\end{aligned}
$$

- Cauchy-Schwartz: $|\mathsf{cov}(V(X),\delta(X))| \leq \sqrt{\mathsf{var}\big(V(X)\big)\,\mathsf{var}\big(\delta(X)\big)}$

$$
1 = |\mathsf{cov}(V(X),\delta(X))| \;\Rightarrow\; \mathsf{var}\big(\delta(X)\big) \geq \frac{1}{I(s)} \qquad \square
$$

Fisher Information and Cramér-Rao Bound: Vector Case

Let $s \in \mathbb{R}^p$ be a (vector) parameter to be estimated from $X \sim f_X(\cdot|s)$

- Score: $V(X) = \nabla_s \log f_X(X|s) = \dfrac{\nabla_s f_X(X|s)}{f_X(X|s)}$ (random vector)
- As in the scalar case: $\mathbb{E}[V(X)] = 0$
- Fisher information matrix:

$$I(s) = -\mathbb{E}\left[\nabla_s^2 \log f_X(X|s)\right]$$
$$= \mathbb{E}\Big[\big(\nabla_s \log f_X(X|s)\big)\big(\nabla_s \log f_X(X|s)\big)^T\Big] = \mathsf{cov}(V(X))$$

- Cramér-Rao bound: for $\delta : \mathcal{X} \to \mathbb{R}^n$,

$$\underbrace{\mathbb{E}[\delta(X)] = s}_{\text{unbiased}} \Rightarrow \mathsf{cov}\big(\delta(X)\big) \succcurlyeq I(s)^{-1},$$

i.e. $A - B$ is positive semi-definite: $u^T A u \geq u^T B u$, for any $u \in \mathbb{R}^n$

Fisher Information and Cramér-Rao Bound: Vector Case

Let $s \in \mathbb{R}^p$ be a (vector) parameter to be estimated from $X \sim f_X(\cdot|s)$

- Cramér-Rao bound: for $\delta : \mathcal{X} \to \mathbb{R}^n$,

$$\underbrace{\mathbb{E}[\delta(X)] = s}_{\text{unbiased}} \Rightarrow \text{cov}\big(\delta(X)\big) \succcurlyeq I(s)^{-1},$$

- If $r = a^T s$ is a linear function of s, $\gamma(X) = a^T \delta(X)$ is its estimate.

$$\text{var}\big(\gamma(X)\big) = a^T \text{cov}\big(\delta(X)\big) a \geq a^T I(s)^{-1} a$$

- For n i.i.d. observations, $X_i \sim f_{X_i}(\cdot|s) = f_X(\cdot|s)$, for $i = 1, ..., n$,

$$\log f_{X_1,...,X_n}(x_1, ..., x_n|s) = \log \prod_{i=1}^{n} f_X(x_i|s) = \sum_{i=1}^{n} \log f_X(x_i|a),$$

$$I_{(n)}(s) = \mathbb{E}\Big[\nabla_s^2 \sum_{i=1}^{n} \log f_X(X_i|s)\Big] = \sum_{i=1}^{n} \mathbb{E}\Big[\nabla_s^2 \log f_X(X|s)\Big] = n I_{(1)}(s)$$

Fisher Information: Exponential Families

Canonical exponential family with parameter $\eta \in \mathbb{R}^p$,

$$f_X(x|\eta) = \frac{1}{Z(\eta)}\, h(x)\, \exp\big(\eta^T \phi(x)\big)$$

- Fisher information matrix (notice that $\nabla_z^2\, z^T a = 0$):

$$I(\eta) = -\mathbb{E}\left[\nabla_\eta^2 \log f_X(X|\eta)\right] = \nabla_\eta^2 \log Z(\eta)$$

- ...but, it is easy to show (try it), that

$$\frac{\partial^2}{\partial \eta_i\, \partial \eta_j} \log Z(\eta) = \mathsf{cov}\big(\phi_i(X)\, \phi_j(X)\big),$$

thus

$$I(\eta) = \nabla_\eta^2 \log Z(\eta) = \mathsf{cov}\big(\phi(X)\big)$$

Maximum Likelihood (ML) Estimation

Canonical exponential family with parameter $\eta \in \mathbb{R}^n$,

$$f_X(x|\eta) = \frac{1}{Z(\eta)}\, h(x)\, \exp\big(\eta^T \phi(x)\big)$$

- Maximum likelihood estimate:

$$\widehat{\eta}(x) = \delta(x) = \arg\max_{\eta'} f_X(x|\eta') = \arg\max_{\eta'} \big[-\log Z(\eta') + (\eta')^T \phi(x)\big]$$

- ...thus $\quad \widehat{\eta}(x) = \text{solution}_{\eta'} \big[\nabla \log Z(\eta') = \phi(x)\big]$

$$= \text{solution}_{\eta'} \big[\mathbb{E}[\phi(X)|\eta'] = \phi(x)\big]$$

- Example: $f_X(x) = \mathcal{N}(x; \eta, 1) \propto \exp(\eta\, x)$
  - $\phi(x) = x$
  - $\mathbb{E}[\phi(X)|\eta'] = \eta'$
  - $\widehat{\eta}(x) = \text{solution}_{\eta'}\, (\eta' = x) = x$

Maximum Likelihood (ML) Estimation: Consistency

Canonical exponential family with n i.i.d. observations $X = (X_1, ..., X_n)$

$$\log f_X(x|\eta) = -n \log Z(\eta) + \eta^T \sum_{i=1}^{n} \phi(x_i)$$

- Maximum likelihood estimate:

$$\widehat{\eta}(x) = \text{solution}_{\eta'} \left(\mathbb{E}[\phi(X)|\eta'] = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \right) \qquad \text{(moment matching)}$$

- Weak law of large numbers:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) = \mathbb{E}[\phi(X)|\eta]$$

- Consistency:

$$\lim_{n \to \infty} \widehat{\eta}(x) = \text{solution}_{\eta'} \left(\mathbb{E}[\phi(X)|\eta'] = \mathbb{E}[\phi(X)|\eta] \right) = \eta$$

...also true with more generality, under some additional conditions.

ML Estimation: Asymptotic Normality

Canonical exponential family with n i.i.d. observations $X = (X_1, ..., X_n)$

$$\log f_X(x|\eta) = -n \log Z(\eta) + \eta^T \sum_{i=1}^{n} \phi(x_i)$$

- Asymptotic normality: let $\widehat{\eta}(x)$ be the ML estimation function

$$\lim_{n\to\infty} \sqrt{n}\big(\widehat{\eta}(X) - \eta\big) \stackrel{d}{=} Z \sim \mathcal{N}\left(0, I(\eta)^{-1}\right)$$

- ...thus (true with more generality, under some conditions)

$$\lim_{n\to\infty} \widehat{\eta}(X) = Z \sim \mathcal{N}\left(\eta, \frac{I(\eta)^{-1}}{n}\right)$$

- Convergence in distribution: $Z_1, Z_2, ..., Z_n, ...$ have distribution functions $F_1, F_2, ..., F_n, ...$; Z has distribution function F

$$\lim_{n\to\infty} Z_n \stackrel{d}{=} Z \quad \text{if} \quad \lim_{n\to\infty} F_n(x) = F(x),$$

at every point x where F is continuous.

The Bias-Variance Trade-off

Let $s \in \mathbb{R}$ be a scalar to be estimated, $X \sim f_X(\cdot|s)$, and $\delta : \mathcal{X} \to \mathbb{R}$.

- Expected estimate: $\bar{s} = \mathbb{E}[\delta(X)]$
- Bias-variance decomposition (of the squared error loss)

$$\begin{aligned}
\mathbb{E}\left[(\delta(X) - s)^2\right] &= \mathbb{E}\left[(\delta(X) - \bar{s} + \bar{s} - s)^2\right] \\
&= \mathbb{E}\left[(\delta(X) - \bar{s})^2\right] + \mathbb{E}\left[(s - \bar{s})^2\right] + 2\mathbb{E}\left[(\delta(X) - \bar{s})(\bar{s} - s)\right] \\
&= \mathbb{E}\left[(\delta(X) - \bar{s})^2\right] + (s - \bar{s})^2 + 2(\bar{s} - s)\underbrace{(\mathbb{E}[\delta(X)] - \bar{s})}_{0} \\
&= \underbrace{\mathbb{E}\left[(\delta(X) - \bar{s})^2\right]}_{\text{variance}} + \underbrace{(s - \bar{s})^2}_{\text{(squared) bias}}
\end{aligned}$$

- The minimum expected loss may be achieved by a biased estimator.
- Generalizes trivially to the vector case.

The Bias-Variance Trade-off: Example

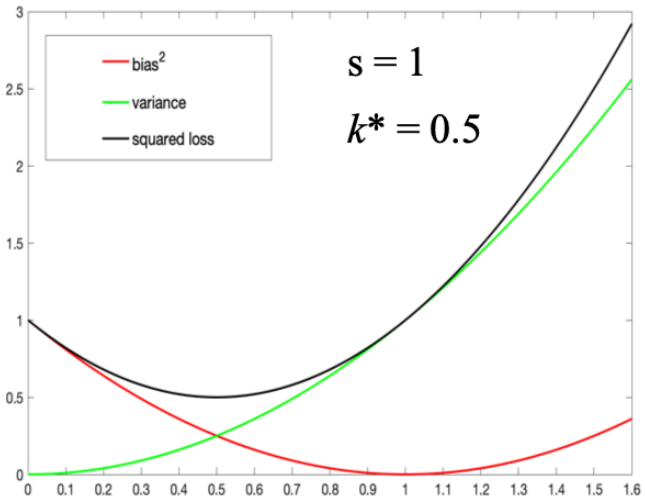
Let $s \in \mathbb{R}$; observations $X \sim \mathcal{N}(s, 1)$; linear estimator: $\delta_k(x) = k\,x$.

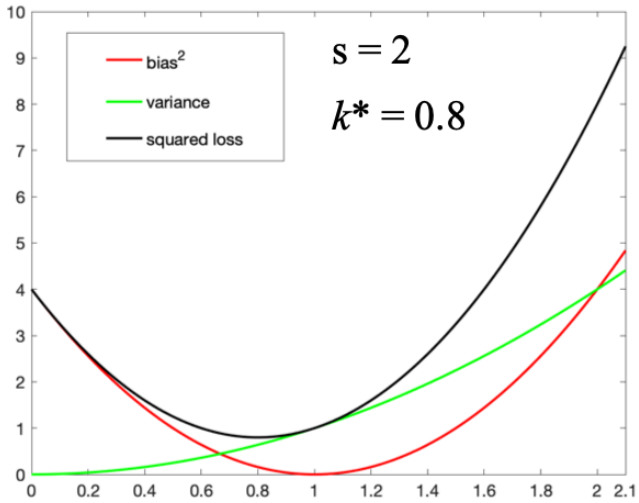
- Bias-variance decomposition:

$$\mathbb{E}\left[(\delta_k(X) - s)^2\right] = \underbrace{k^2}_{\text{variance}} + \underbrace{s^2(k-1)^2}_{\text{bias}^2}$$

(see slide 7)

- Optimal k (minimum squared error loss): $k^* = s^2/(s^2+1)$





- Of course, in practice, s is unknown.

Stein's "Paradox" and James-Stein Estimation

Estimate $\theta \in \mathbb{R}^n$ from $f_X(x|\theta) = \prod_{i=1}^{n} \mathcal{N}(x_i; \theta_i, 1)$, with $n \geq 3$.

- ML estimate (unbiased): $\widehat{\theta}_{\mathsf{ML}}(x) = \delta_{\mathsf{ML}}(x) = x$.
- James-Stein estimate: $\widehat{s}_{\mathsf{JS}}(x) = \delta_{\mathsf{JS}}(x) = x \underbrace{\left(1 - \frac{n-2}{\|x\|_2^2}\right)_+}_{<1}$ (shrinkage)
- Risk of δ_{ML}: $R(\theta, \delta_{\mathsf{ML}}) = \mathbb{E}\left[\|\delta_{\mathsf{ML}}(X) - \theta\|_2^2\right] = \mathbb{E}\left[\|X - \theta\|_2^2\right] = n$
- Risk of δ_{JS}: $R(\theta, \delta_{\mathsf{JS}}) = n - \underbrace{(n-2)^2 \, \mathbb{E}\left[\frac{1}{\|X\|_2^2}\right]}_{>0} < R(\theta, \delta_{\mathsf{ML}})$
- ...the ML estimator, in this problem, is inadmissible, which seems paradoxical, given conditionally independent observations.

James-Stein Estimation: A Bias-Variance View

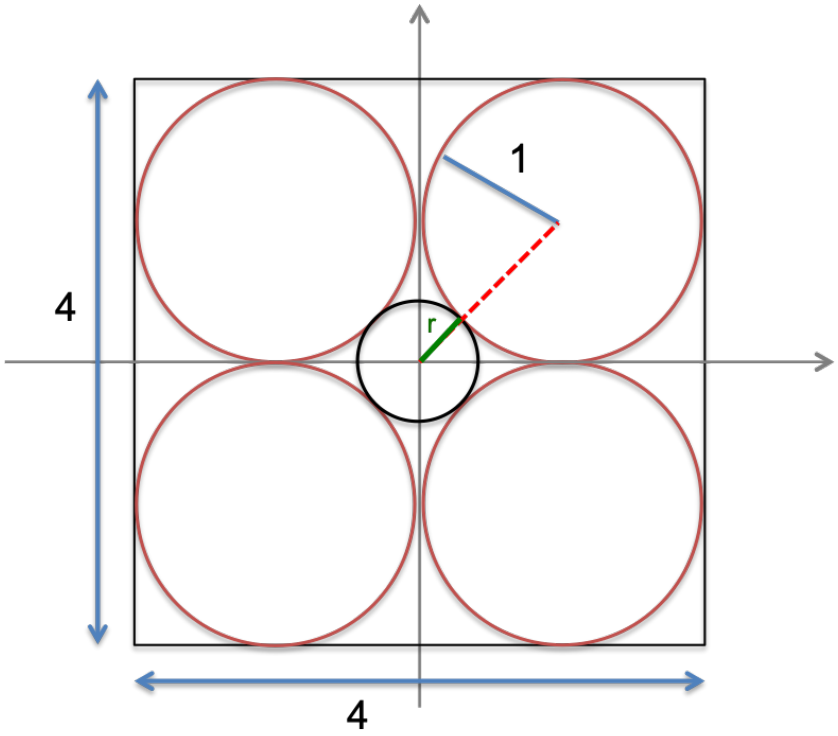
Estimate $\theta \in \mathbb{R}^n$ from $f_X(x|\theta) = \prod_{i=1}^{n} \mathcal{N}(x_i; \theta_i, 1)$, with $n \geq 3$.

- Estimators of the form $\delta_\lambda(x) = \lambda\, x$, for $\lambda \in \mathbb{R}_+$
- Bias-variance decomposition:

$$\begin{aligned} \mathbb{E}\left[\|\delta_\lambda(X) - \theta\|^2\right] &= \overbrace{\mathbb{E}\big[\|\delta_\lambda(X) - \underbrace{\lambda\mathbb{E}[X]}_{\mathbb{E}[\delta_\lambda(X)]}\|^2\big]}^{\text{variance}} + \overbrace{\|\theta - \lambda\mathbb{E}[X]\|^2}^{\text{(squared) bias}} \\ &= \lambda^2 n + (\lambda - 1)^2 \|\theta\|^2 \end{aligned}$$

- Optimal choice (not usable, θ is unknown): $\lambda^* = \dfrac{\|\theta\|^2}{\|\theta\|^2 + n} \in]0, 1[$
- The James-Stein factor $(1 - (n-2)/\|x\|^2)_+$ is (approximately) an unbiased estimate of λ^*.
- Only for $\|\theta\| \to \infty$, does $\lambda \to 1$. For finite $\|\theta\|$: shrinking is good.

Dessert: High-Dimensional Intuition Failure



- In $\mathbb{R}^2$:
  - 4 unit-radius circles inside $[-2, 2]^2$
  - circle of radius r centred at the origin, confined by the unit-radius circle
  - distance from origin to centers: $\sqrt{2}$
  - $r = \sqrt{2} - 1$
- In $\mathbb{R}^d$:
  - 2^d unit-radius spheres inside $[-2, 2]^d$
  - sphere of radius r centred at the origin, confined by the unit-radius spheres
  - $r = \sqrt{d} - 1$
  - for $d \geq 10$, we have $r > 2$

- J. Michael Steele, *The Cauchy-Schwarz Master Class*, Cambridge, 2004.

Recommended Reading

- K. Murphy, "Machine Learning: A Probabilistic Perspective", MIT Press, 2012 (chapter 6).
- L. Wasserman, "All of Statistics: A Concise Course in Statistical Inference", Springer, 2004.
- M. Figueiredo, "Lectures Notes on Bayesian Estimation and Classification", unpublished manuscript, available at `https://fenix.tecnico.ulisboa.pt/disciplinas/AEsta-4/2014-2015/2-semestre/lecture-notes`