Nonstandard analysis and statistical decision theory

David Schrittesser
Harbin Institute of Technology,
University of Toronto

SETTOP 2022

All results are joint work with
Dan M. Roy
and
Haosui (Kevin) Duanmu.

1. Want to measure an empirical quantity ξ
2. Make n (imprecise) measurements, obtaining $x_1, \ldots, x_n$.
3. Give an estimate of ξ, as a function of $\vec{x} = (x_1, \ldots, x_n)$, e.g.:

$$\hat{\xi}(\vec{x}) = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

or

$$\hat{\xi}(\vec{x}) = \frac{\min\{x_1, \ldots, x_n\} + \max\{x_1, \ldots, x_n\}}{2}$$

But which one of these, and why?

And why not completely other estimates?

Perhaps we want to estimate the precision of the measurement.

E.g., by

$$\hat{\sigma}(\vec{x}) = \frac{\sum_{i=1}^{n}(\bar{x} - x_i)^2}{n-1}$$

or, for some $a > 0$, by

$$\hat{\sigma}_a(\vec{x}) = \frac{\sum_{i=1}^{n}(\bar{x} - x_i)^2}{a}$$

e.g., with $a = n + 1$.

But again: Which one of these? Why?

And why not completely other estimates?

To compare estimators, be inspired by... game theory!

Player I (Nature) chooses $\theta \in \Theta$

- Θ ...... Parameterspace (possible states of nature)

Player II (Statistician) chooses $\delta \colon \mathbb{X} \to \mathbb{A}$ from $\mathcal{D}$

- $\mathbb{X}$ ...... Samplespace (possible measurement outcomes)
- $\mathbb{A}$ ...... Actionspace (possible estimates, or accept/reject H_0)
- $\mathcal{D}$ ...... decision procedures available to Statistician

Outcome of the game: Player II suffers a loss of

$$r(\theta, \delta)$$

Where do we get $r(\theta, \delta)$ from?

1. Fix a family of measures $(P_\theta)_{\theta \in \Theta}$ on $\mathbb{X}$.
   Assume that under the condition that Nature chooses θ, the probability of measuring $x \in B$ is given as:

$$P(x \in B \mid \theta) = P_\theta(B)$$

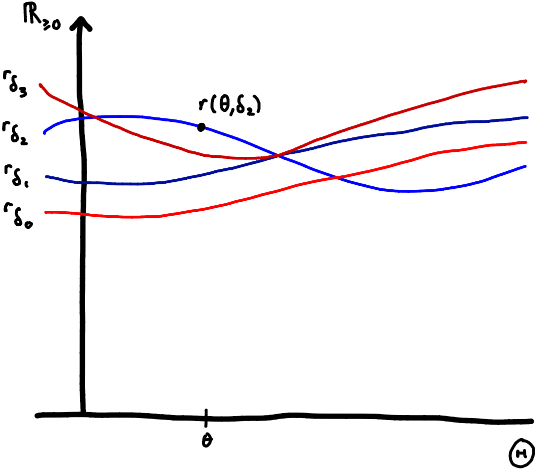
2. Fix a loss function

$$(\theta, \hat{\theta}) \mapsto \ell(\theta, \hat{\theta}) \in \mathbb{R}_{>0}$$

3. For each $\delta \colon \mathbb{X} \to \mathbb{A}$, define its risk function as its expected loss

$$\begin{aligned} r^\delta(\theta) \equiv r(\theta, \delta) &:= \int_{\mathbb{X}} \ell\big(\theta, \delta(x)\big) P_\theta(\mathrm{d}x) \\ &= \mathbb{E}_\theta \, \ell(\theta, x) \end{aligned}$$

Figure: Some risk function in ${}^{\Theta}\mathbb{R}$



Example: Normal location

Let $\mathbb{X} = (\mathbb{R}^d)^n$, i.e., we take n samples $x_1, \dots, x_n$ from $\mathbb{R}^d$.

$$P(x \in B \mid \mu, \sigma) = P_{\mu,\sigma}(B) \propto \prod_{j=1}^{n} \frac{1}{\sigma} \int_B \exp\left(-\frac{\|\mu - x_j\|^2}{2\sigma^2}\right)$$

Let the loss function be given by

$$\ell(\mu, \hat{\mu}) = \|\mu - \hat{\mu}\|^2$$

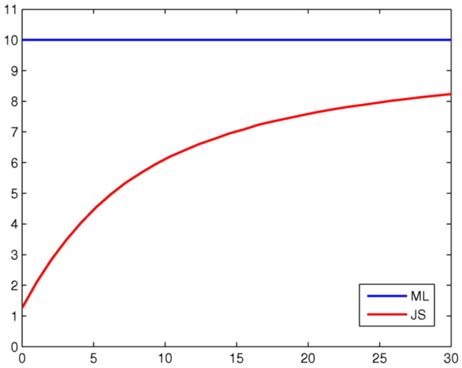
Consider

$$\hat{\mu}_{\mathrm{ML}}(x) = \bar{x}$$

If $d > 2$, also consider the James-Stein estimator,

$$\hat{\mu}_{\mathrm{JS}}(x) = \left(1 - \frac{(d-2)\bar{s}}{(n+1)\|x\|^2}\right)\bar{x}, \text{ with } \bar{s} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Surpisingly, $\hat{\mu}_{\mathsf{JS}}$ outperforms $\hat{\mu}_{\mathsf{ML}}$:



Note:

- $\hat{\mu}_{\mathsf{JS}}$ is biased: $\mathbb{E}_{\mu}\,\hat{\mu}_{\mathsf{JS}}(x) \neq \mu$.
- among the unbiased estimators, $\hat{\mu}_{\mathsf{ML}}$ has uniform minimum risk.

To each $\delta \in \mathcal{D}$ corresponds a point in risk space

$$^{\Theta}\mathbb{R},$$

namely

$$r^{\delta} = \text{the element of } {}^{\Theta}\mathbb{R} \text{ given by } \theta \mapsto r(\theta, \delta).$$

We call

$$R^{\mathcal{D}} = \{r^{\delta} \mid \delta \in \mathcal{D}\}$$

the risk set corresponding to $\mathcal{D}$ and r.

Another important notion is equivalence in risk,

$$\delta \sim \delta' \overset{\text{def}}{\iff} r^{\delta} = r^{\delta'}.$$

In some contexts, we identify rules which are equivalent in risk.

Admissibility

Decision rules are partially ordered by

$$\delta' \preceq \delta \iff (\forall\theta \in \Theta)\ r(\theta, \delta') \leq r(\theta, \delta).$$

The strict part of this partial order is domination,

$$\begin{aligned} \delta' \prec \delta &\iff \delta' \preceq \delta \wedge \delta \not\preceq \delta' \\ &\iff \delta' \preceq \delta \wedge (\exists\theta \in \Theta)\ r(\theta, \delta') < r(\theta, \delta). \end{aligned}$$

δ is admissible among $\mathcal{D} \iff \neg\exists\delta' \in \mathcal{D}$ such that $\delta' \prec \delta$.

- Necessary but very insufficient for optimality: Constant estimators are often admissible!
- Admissibility of some interesting procedures, e.g., the so-called Graybill-Deal estimator, is an open problem

Admissibility is a frequentist notion: The state of nature θ is assumed to be unknown, but fixed.

Bayesian methods take a different approach:

Assume θ is itself a random variable, i.e., its behaviour is given by a prior probability distribution

$$\pi \in \mathcal{P}_1(\Theta),$$
$$\pi(B) = \text{probability that } \theta \in B.$$

Define the Bayes risk of δ under π as

$$r(\pi, \delta) = \int r(\theta, \delta)\, \pi(\mathrm{d}\theta)$$

This induces a total preordering on $\mathcal{D}$.

A mimimum is called a Bayes rule w.r.t. π.

Lemma

Suppose δ is a Bayes rule w.r.t. π and

$$\pi(U) > 0 \textit{ for every non-empty open } U$$

and that for all $\delta \in \mathcal{D}$, $\theta \mapsto r(\theta, \delta)$ is continuous. Then δ is admissible.

Proof.

Suppose otherwise that $\delta' \prec \delta$. There is $U \neq \emptyset$ open such that $(\forall \theta \in U)\, r(\theta, \delta') < r(\theta, \delta)$. Since $\pi(U) > 0$,

$$\begin{aligned} r(\pi, \delta') &= \int_U r(\theta, \delta')\, \pi(\mathrm{d}\theta) + \int_{\Theta \setminus U} r(\theta, \delta')\, \pi(\mathrm{d}\theta) \\ &< \int_U r(\theta, \delta)\, \pi(\mathrm{d}\theta) + \int_{\Theta \setminus U} r(\theta, \delta)\, \pi(\mathrm{d}\theta) = r(\pi, \delta). \quad \square \end{aligned}$$

Corollary

Any decision rule δ which is Bayes with respect to a prior π such that

$$(\forall \theta \in \Theta)\ \pi(\{\theta\}) > 0$$

is admissible.

If Θ is finite, there is also an implication from admissible to Bayes:

Theorem (Wald?)

If Θ is finite, $R^{\mathcal{D}}$ is convex, and $\delta_0 \in \mathcal{D}$ is admissible, there is a prior $\pi \in \mathcal{P}_1(\Theta)$ such that δ_0 is π-Bayes.

As in our example, square error is a commonly used loss function:

$$\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2.$$

If $\mathbb{A}$ is a convex set, this function is convex in the action:

For $\lambda_i \in [0, 1]$, $a_i \in \mathbb{A}$ $(i < n)$ with $\sum_i \lambda_i = 1$,

$$\ell(\theta, \sum_i \lambda_i a_i) \leq \sum_i \lambda_i \ell(\theta, a_i)$$

Then, $\mathcal{D}$ with

$$\left(\sum_i^{\mathcal{D}} \lambda_i \delta_i \right)(x) = \sum_i^{\mathbb{A}} \lambda_i \delta_i(x)$$

becomes a convex set and $r(\theta, \cdot)$ is convex for each θ.

One can cover a wider class of problems through "randomization":

Instead of considering procedures

$$\delta \colon \mathbb{X} \to \mathbb{A}$$

allow

$$\delta \colon \mathbb{X} \to \mathcal{P}_1(\mathbb{A})$$

with the interpretation that the statistician takes a random action $a \in \mathbb{A}$ distributed as $\delta(x)$.

The risk is (re)defined as the expected loss and becomes linear in δ:

$$\begin{aligned} r(\theta, \delta) &= \int_{\mathbb{X}} \ell(x, a)\ \delta(x)(\mathrm{d}a)\ P_\theta(\mathrm{d}x) \\ &= \mathbb{E}_\theta\, \ell\big(x, \delta(x)\big) \end{aligned}$$

From now on, assume $\mathcal{D}$ is convex and $r(\theta, \delta)$ is linear in δ.

Decision Theoretic Framework

Components of a statistical decision problem:

- parameterspace Θ,
- sample space $\mathbb{X}$,
- the model $(P_\theta)_{\theta \in \Theta}$,
- action space $\mathbb{A}$,
- loss function $\ell\colon \Theta \times \mathbb{A} \ \to [0, \infty)$,
- The set of randomized decision rules $\mathcal{D}$.

Given the unknown state of nature $\theta \in \Theta$, $x \in \mathbb{X}$ is drawn from P_θ. Statistician observes x, then selects an action $a \in \mathbb{A}$ according to $\delta(x)$ and suffers the loss $\ell(\theta, a)$.

Goal: Find $\delta\colon \mathbb{X} \to \mathcal{P}_1(\mathbb{A})$ which minimizes (in a specified sense) the expected loss, a.k.a. the risk

$$r(\theta, \delta) = \mathbb{E}_\theta \, \ell\big(x, \delta(x)\big)$$

Connections between frequentist and Bayesian optimality

An interpretation of the (frequentist) notion of admissibility in a Bayesian framework has been a long-standing goal.

A rule δ is admissible when...

- δ has minimal Bayes risk w.r.t. a an "everywhere positive" prior π, provided risk functions are continuous or Θ countable
- δ is the unique (up to $\sim$) Bayes rules for some π

Admissible rules are Bayes provided...

- Θ is finite (Wald)
- under compactness and continuity conditions on some or all of Θ, $\mathbb{X}$, $\mathbb{A}$, $\mathcal{D}$, r, ℓ (Wald, Berger)

More partial equivalences using: limit of Bayes, generalized Bayes, under technical conditions (Wald, LeCam, Brown, Stone, Berger, Srinivasan)

Indeed, there are rules which are admissible but *not* Bayes:

Example

In the multivariate normal location problem in 2 dimensions under mean square error, the "usual" estimator is admissible but *not* Bayes.

We identify an exact equivalence between frequentist admissibility and Bayes optimality once we allow priors to assign infinitesimal mass to certain sets.

A precursor:

Theorem (Duanmu-Roy, 2017)

A decision rule is extended admissible if and only if it is non-standard Bayes.

(For this talk, you don't need to know what "extended admissible" and "non-standard Bayes" are.)

Nonstandard Decision Theory

We work in a superstructure:

$$V(\mathbb{R}) := V_\omega(\mathbb{R}) = \bigcup_{n\in\mathbb{N}} V_n(\mathbb{R}) = \mathbb{R} \cup \mathcal{P}(\mathbb{R}) \cup \dots,$$
$$^*V(\mathbb{R}) := V(\mathbb{R})^I/\mathcal{U}$$

where $\mathcal{U}$ is an ultrafilter on a set I.

Nonstandard people call the elementary embedding the star map,

$$^*(\cdot)\colon V(\mathbb{R}) \to {}^*V(\mathbb{R}),$$
$$x \mapsto {}^*x$$

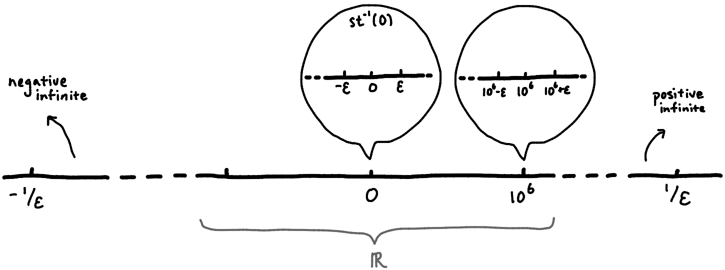
We can ask that $^*V(\mathbb{R})$ is saturated:
If $\{\phi_\xi(x) \mid \xi < \theta\}$ is finitely satisfiable by elements of $V_n(\mathbb{R})$, then

$$(\exists x \in {}^*V_n(\mathbb{R}))\ (\forall \xi < \theta)\ {}^*\phi_\xi(x).$$

Thus, in ${}^*\mathbb{R}$, there are infinitesimals:

$$(\exists \varepsilon \in {}^*\mathbb{R})\ 0 < r \wedge (\forall n \in \mathbb{N})\ \varepsilon < \frac{1}{n}$$

Figure: The hyperreals ${}^*\mathbb{R}$



Hyperpriors

A hyperprior is an element Π of ${}^*\mathcal{P}_1(\Theta)$.
That is, a ${}^*\sigma$-additive map

$$\Pi\colon {}^*\mathcal{P}(\Theta) \to {}^*[0,1]$$

with $\Pi({}^*\Theta) = 1$. The set

$${}^*[0,1]$$

consists of reals of the form

$$r = \underbrace{r'}_{=\mathrm{st}(r)} + \varepsilon$$

where $r' \in [0,1]$ and $\varepsilon \in {}^*\mathbb{R}$ is infinitesimal.

In contrast to an ordinary prior, Π can assign infinitesimal weight!

An Example: multivariate normal location

Consider estimating the mean of an n-dimensional multivariate normal distribution, given just one sample $\vec{x}$.

Let K be infinite and take the non-standard prior

$$\Pi_K(\mathrm{d}\vec{\mu}) \propto \frac{1}{K} \exp(-\frac{1}{2K^2}\mu^2)$$

The prior density is near constant on $\mathbb{R}^n$.
The corresponding Bayes rule is:

$$\delta_{\Pi_K}(\vec{x}) = \frac{K^2}{K^2+1}\vec{x}$$

but for $n = 1$, the "usual" estimate

$$\delta(\vec{x}) = \vec{x}$$

has minimal risk under Π_K among the standard rules.

Characterization of admissibility

Allowing priors with infinitesimals, we can give a Bayesian interpretation of admissibility.

Theorem (DR**S**-21)

A decision rule δ_0 *is admissible among* $\mathcal{D}$ *if and only if there exists a hyperprior* Π *on* $^*\Theta$ *such that*

1. $^*r(^*\delta_0, \Pi) \leq {}^*r(^*\delta, \Pi)$ *for all* $\delta \in \mathcal{D}$ *and*
2. $\Pi(^*\theta) > 0$ *for all* $\theta \in \Theta$*.*

Some ideas from the proof

Recall:

Theorem (Wald?)

If Θ is finite and $\delta_0 \in \mathcal{D}$ is admissible, there is a prior $\pi \in \mathcal{P}_1(\Theta)$ such that δ_0 is π-Bayes.

Theorem (Blackwell-Girshick?)

Suppose Θ is finite, $\mathcal{D}$ is the convex hull of finitely many points, and $\delta_0 \in \mathcal{D}$ is admissible. Then there is a prior $\pi \in \mathcal{P}_1(\Theta)$ such that δ_0 is π-Bayes and $\pi(\theta) > 0$ for all $\theta \in \Theta$.

Lemma

Suppose δ_0 is admissible and $\mathcal{D}_0 \subseteq \mathcal{D}$ is finite.
Then there is a finite set $\Theta_0 \subseteq \Theta$ such that for each $\delta \in \mathrm{conv}(\mathcal{D}_0) \setminus \{\delta_0\}$, there is $\theta \in \Theta_0$ such that $r(\theta, \delta_0) < r(\theta, \delta)$.

Fact

Given $X \in V(\mathbb{R})$ we can find hyperfinite $\tilde{X} \in {}^*V(\mathbb{R})$ such that

$$\{{}^*x \mid x \in X\} \subseteq \tilde{X} \subseteq {}^*X.$$

Proof.

The following gives a finitely satisfiable set of sentences:

$$\phi_x(Y) := Y \text{ is finite and } x \in Y \quad (x \in X)$$

By saturation there exists $\tilde{X}$ satisfying

$$(\forall x \in X)\ {}^*\phi_x(\tilde{X})$$

i.e., $\tilde{X}$ is hyperfinite and $\{{}^*x \mid x \in X\} \subseteq \tilde{X}$. $\square$

Theorem
If δ_0 is admissible among $\mathcal{D}$, there exists a hyperprior Π on ${}^*\Theta$ such that

1. ${}^*r({}^*\delta_0, \Pi) \leq {}^*r({}^*\delta, \Pi)$ *for all* $\delta \in \mathcal{D}$ *and*
2. $\Pi({}^*\theta) > 0$ *for all* $\theta \in \Theta$.

Proof.
Find hyperfinite $\tilde{\Theta}$ and $\tilde{\mathcal{D}}$ s.t.

$$\Theta \subseteq \tilde{\Theta} \subseteq {}^*\Theta,$$
$$\{{}^*\delta \mid \delta \in X\} \subseteq \tilde{\mathcal{D}} \subseteq {}^*\mathcal{D}.$$

By transfer, ${}^*\delta_0$ is admissible among ${}^*\mathcal{D} \supseteq {}^*\mathrm{conv}(\tilde{\mathcal{D}})$.
By the transfer of a previous Lemma, we can assume admissibility of ${}^*\delta_0$ among ${}^*\mathrm{conv}(\tilde{\mathcal{D}})$ is witnessed on $\tilde{\Theta}$.
By *Blackwell-Girshick, there exists a hyperprior Π as required. $\square$

Blyth's method

Theorem

Suppose $\Theta \subseteq \mathbb{R}^n$ is open, procedures with continuous risk functions form a complete class ($\equiv$ every discontinuous procedure is dominated by a continuous one), and δ_0 has continous risk.

Then δ_0 is admissible if there is a sequence $\pi_0, \pi_1, \dots$ of measures such that

- $r(\pi_n, \delta_0) < \infty$ *for all* $n \in \mathbb{N}$,
- *For any non-empty open* $O \subseteq \Theta$,

$$\lim_{n\to\infty} \frac{r(\pi_n, \delta_0) - r(\pi_n, \delta^{\pi_n})}{\pi_n(O)} = 0$$

An Application: Nonstandard Blyth

Theorem (DR**S**22+)

δ_0 is admissible iff there exists

- $\Pi \in {}^*\big(\mathcal{P}_1(\Theta)\big)$
- $\tilde{\rho} \in {}^*\mathbb{R}$ *with* $\tilde{\rho} > 0$

such that

1. $\tilde{\rho} \le \Pi(\theta)$ *for all* $\theta \in \Theta$,
2. $$\frac{{}^*r(\Pi, {}^*\delta_0) - \inf_{\delta \in \mathcal{D}} {}^*r(\Pi, {}^*\delta)}{\tilde{\rho}} \approx 0$$

An Application (estimating a common normal location)

Suppose we have two groups of random variables,

$$X_{i,1}, \ldots, X_{i,n} \quad (i = 0, 1)$$

where each group is i.i.d. as follows:

$$X_{0,j} \sim \mathcal{N}(\mu, \sigma_0), \qquad X_{1,j} \sim \mathcal{N}(\mu, \sigma_1)$$

If σ_0, σ_1 are known,

$$\hat{\mu}(x) = \frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2}\bar{x}_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}\bar{x}_1$$

is a reasonable estimator.

For unknown σ_0, σ_1, Graybill-Deal (1951) suggested

$$\hat{\mu}_{\mathrm{GD}}(x) = \frac{s_1^2}{s_0^2 + s_1^2}\bar{x}_0 + \frac{s_0^2}{s_0^2 + s_1^2}\bar{x}_1$$

where

$$s_i^2 = \frac{\sum_j (\bar{x}_i - x_{i,j})^2}{n-1}$$

- Known to be extended admissible among scale and location invariant estimators
- Not known to be admissible (among all estimators)

Let $\mathcal{C}$ be the class of all estimators of the form

$$\hat{\mu}(x) = \bar{x}_0 + (\bar{x}_1 - \bar{x}_0) \cdot \hat{\phi}(s_1^2, s_2^2)$$

for an arbitrary function $\hat{\phi}$.

Note

The Graybill-Deal estimator itself is of this form:

$$\hat{\mu}_{\mathrm{GD}}(x) = \bar{x}_0 + (\bar{x}_1 - \bar{x}_0) \cdot \frac{s_1^2}{s_1^2 + s_1^2}$$

Theorem

The Graybill-Deal estimator $\hat{\mu}_{GD}$ is admissible among $\mathcal{C}$.

Thank You!

Find our paper at `https://arxiv.org/`

Lemma

Suppose δ_0 is admissible, $\mathcal{D}_0 \subseteq \mathcal{D}$ is finite, and $\delta_0 \notin \text{conv}(\mathcal{D}_0)$. Then there is a finite set $\Theta_0 \subseteq \Theta$ such that for each $\delta \in \text{conv}(\mathcal{D}_0)$, there is $\theta \in \Theta_0$ such that $r(\theta, \delta_0) < r(\theta, \delta)$.

Proof.

Otherwise, for every finite $\Theta_0 \subseteq \Theta$ there is $\delta \in \text{conv}(\mathcal{D}_0)$ such that $r^\delta \restriction \Theta_0 \leq r^{\delta_0} \restriction \Theta_0$.
By saturation, there exists $\Delta \in {}^*\text{conv}(\mathcal{D}_0)$ such that ${}^*r^\Delta \restriction \Theta \leq r^{\delta_0}$.
Write

$$\Delta = \Lambda_0 {}^*\delta_0 + \ldots + \Lambda_n {}^*\delta_n$$

But then letting

$$\delta := \mathrm{st}(\Lambda_0)\delta_0 + \ldots + \mathrm{st}(\Lambda_n)\delta_n$$

we have $r^\delta \leq r^{\delta_0}$. By admissibility, $r^\delta = r^{\delta_0}$. Hence $r^{\delta_0} \in \text{conv}(R^{\mathcal{D}_0})$, contradiction. □

Lemma

Suppose δ_0 is admissible and $\mathcal{D}_0 \subseteq \mathcal{D}$ is finite.
Then there is a finite set $\Theta_0 \subseteq \Theta$ such that for each $\delta \in \mathrm{conv}(\mathcal{D}_0) \setminus \{\delta_0\}$, there is $\theta \in \Theta_0$ such that $r(\theta, \delta_0) < r(\theta, \delta)$.

Proof.

Can assume $\delta_0 \in \mathrm{conv}(\mathcal{D}_0)$.

1. Decompose $\mathrm{conv}(\mathcal{D}_0)$ into convex sets $C_0, \ldots, C_m$ each having δ_0 as an extreme point.
2. Can assume that δ_0 is an extreme point of $\mathrm{conv}(\mathcal{D}_0)$. Let $\mathcal{D}_0'$ be the set of extreme points of $\mathrm{conv}(\mathcal{D}_0)$, excluding δ_0. Choose Θ_0 as in the previous lemma.
3. For every $\delta \in \mathrm{conv}(\mathcal{D}_0)$ is a convex combination

$$r^{\delta} = \lambda r^{\delta_0} + \lambda' r^{\delta'}$$

with $\delta' \in \mathcal{D}_0'$. For some $\theta \in \Theta_0$, $r^{\delta'}(\theta) > r^{\delta_0}(\theta)$ and so $r^{\delta}(\theta) = \lambda r^{\delta_0}(\theta) + \lambda' r^{\delta'}(\theta) > r^{\delta_0}(\theta)$. □