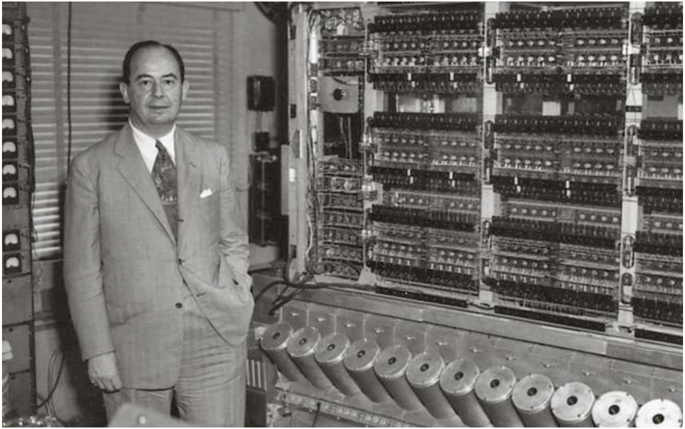
The Rao-Blackwell Theorem

Daniel G. Alabi

Columbia University

August 26, 2022

Outline



John Von Neumann: Hungarian-American Mathematician.

Why (and How) Things Work

In Honor of David Blackwell

Daniel Alabi

LINKS

About David Blackwell
My Personal Website
My Github

RECENT POSTS

PhD Defended!
Kemeny Rank Aggregation
Differential Privacy
Introduction to Sentiment Analysis: A Tale of Distinct Classes

An Application of Linear Programming in Game Theory

*I took the Combinatorial Optimization class at AIT Budapest (Aquincum Institute of Technology) with **David Szeszler**, a Professor at the Budapest University of Technology and Economics. We touched on some Graph Theory, Linear Programming, Integer Programming, the Assignment Problem, and the Hungarian method. My favorite class in the course was focused on applying Linear Programming in Game Theory. I'll summarize the most important aspects of that class in this blog post. I hope this piques your interest in Game Theory (and in attending* AIT).

Basics of Linear Programming

First, I want to touch on some topics in Linear Programming for those who don't know much about setting up a linear program (which is basically a system of linear inequalities with a maximization function or a minimization function). You can skip this section if you are confident about the subject.

Linear Programming is basically a field of mathematics that has to do with determining the optimum value in a *feasible region*. In determining the optimum value, one of two questions can be asked: find the minimum point/region or find the maximum point/region. The feasible region in a linear program is determined by a set of linear inequalities. For a feasible region to even exist, the set of linear inequalities must be solvable.

A typical linear program is given in this form: $max\{cx : Ax \leq b\}$. c is a row vector of dimension n. A is an $m \times n$ matrix called the *incidence matrix*. x is a column vector of dimension n. This is called the **primal program**. The primal program is used to solve *maximization* problems. The dual of this primal program is of the form $min\{yb : yA = c, y \geq 0\}$. b, A, c are the same as previously defined. y is a row vector of dimension m. This is called the **dual program**. The dual is just a derivation of the primal program that is used to solve *minimization* problems.

Having introduced primal and dual programs, the next important theory in line is the **duality theorem**. The duality theorem states that $max\{cx : Ax \leq b\}$ = $min\{yb : yA = c, y \geq 0\}$. In other words, the maximum of the primal program is equal to the minimum of the dual program (provided that the primal program is solvable and bounded from above). Using this "tool", every minimization problem can be converted to a maximization problem and vice versa (as long as the initial problem involves a system of linear inequalities that can be set up as a linear program with a finite amount of linear constraints and one *objective function*).

Posted December 23, 2012! Trending on HackerNews.

Question about Payoffs

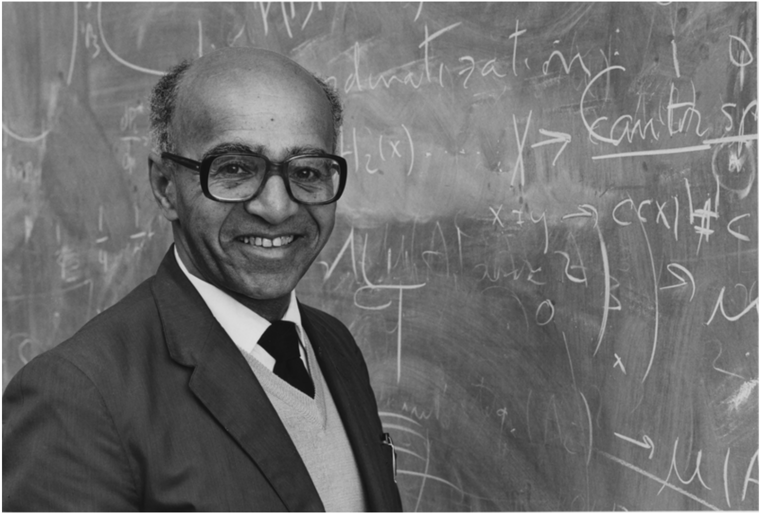
In many cases, it seems that you are calculating the result of the game based on the input of one player only.

(of course, with high probability, this just means I did not get it)

Question that led me to discover Blackwell's Approachability Theorem.

Von Neumann's Theorem (1928) is about scalar utility between two players.
Blackwell asked: what can we achieve with a vector-valued payoff?

Do Black People Ask Questions Like This?



Discovered he looked like me!

- He was interviewed by statistician Jerzy Neyman, who supported his appointment.
- The head of the mathematics department, Griffith C. Evans, supported his appointment, at first, and convinced the university president Robert Sproul.

Shortly after this conversation—before a formal offer could be made to Blackwell—Neyman learned that the wife of a mathematics professor, born and bred in the south, had said that she could not invite a Negro to her house or attend a departmental function at which one was present.

From "Neyman" by Constance Reid.

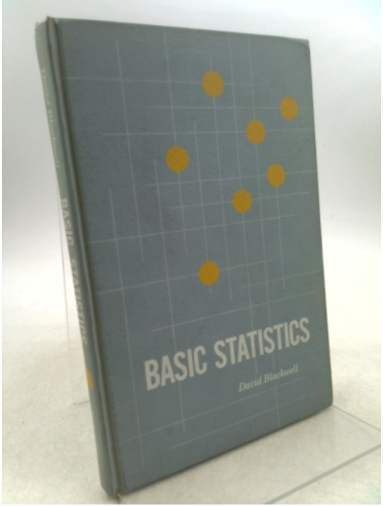
1. Am I *willing* and *ready* to be in the academic system?
2. Volunteered at the Bronx Writing Academy (in New York) and decided that I was.

Bronx Writing Academy.

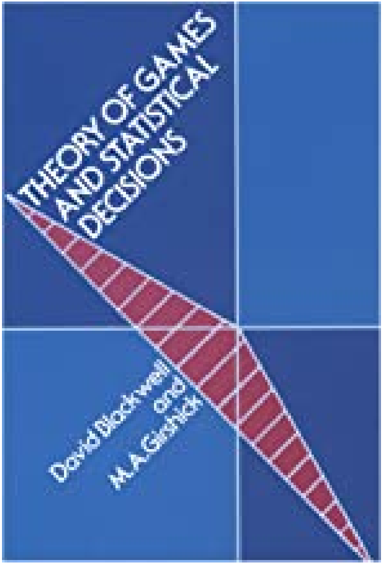
Some of Blackwell's Contributions

1. Inspiring many Black mathematicians/statisticians and computer scientists.
2. Blackwell's Approachability Theorem.
3. **This Talk**: Rao-Blackwell Theorem.
4. Blackwell's Theorem for Contraction Mappings and Dynamic Programming.
5. ...
6. Style of research:
   understanding $\gg$ *publishing for publishing/fame sake*.

1969 Book by David Blackwell



1954 Book by David Blackwell



$\mathcal{D}_\theta$ is a distribution. Let $X \sim \mathcal{D}_\theta$. Then $T(X)$ is a **sufficient statistic** if the conditional distribution of X given $T(X) = T(x)$ does not depend on θ.

Intuition: Cannot gain any more information from the sample to estimate the value of the (unknown) parameter θ.

For any arbitrary distribution, the sample median is not sufficient for the population mean. The sample mean is though.
e.g., consider the sample $1, 2, 3, 4, 100, 100, 100$.

We will show, via the Rao-Blackwell Theorem, that we can rely on sufficient statistics (or in general, **Rao-Blackwellization**) to get *smaller* MSE (Mean-Squared-Error).

How to Know if a Statistic is Sufficient?

Fisher-Neyman Factorization characterizes a sufficient statistic. If $f_\theta(x)$ is the PDF. Then T is **sufficient** for θ iff there exists nonnegative functions g and h can be found such that

$$f_\theta(x) = h(x) \cdot g_\theta(T(x)).$$

$X_1, \ldots, X_n \sim \text{Bern}(p)$.
Joint PDF: $\mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$.
Assuming independence, we get

$$\begin{align}
P(X = x) &= p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \cdots p^{x_n}(1-p)^{1-x_n} \tag{1}\\
&= p^{\sum x_i}(1-p)^{n-\sum x_i} \tag{2}\\
&= p^{T(x)}(1-p)^{n-T(x)}. \tag{3}
\end{align}$$

Mean-Squared-Error (MSE)

Let $\hat{\theta}$ be an estimator of θ.

$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = (\mathrm{Bias}(\hat{\theta}))^2 + \mathrm{Var}(\hat{\theta})$.

$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$.

$\mathrm{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$.

Two Estimators: Which is Better?

$X_1, \ldots, X_n \sim \text{Bern}(p)$.
Define:

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. $X_{15} = \frac{1}{5}(X_1 + 4X_2)$.

Expected Value: Sufficient vs. Non-Sufficient

$X_1, \dots, X_n \sim \text{Bern}(p)$.
Define:

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. $X_{15} = \frac{1}{5}(X_1 + 4X_2)$.

$\mathbb{E}[\bar{X}] = \mathbb{E}[X_{15}] = p$.

Variance: Sufficient vs. Non-Sufficient

$X_1, \ldots, X_n \sim \mathrm{Bern}(p)$.
Define:

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. $X_{15} = \frac{1}{5}(X_1 + 4X_2)$.

$\mathrm{Var}[\bar{X}] = \frac{p(1-p)}{n}$.

$\mathrm{Var}[X_{15}] = \frac{17}{25}p(1-p)$.

We see that

$$\frac{\mathrm{Var}(\bar{X})}{\mathrm{Var}(X_{15})} \to 0.$$

In this case $\bar{X}$ (the MLE) is a better estimator than X_{15}. Is this general?

In this case $\bar{X}$ (the MLE) is a better estimator than X_{15}. Is this general? **No**.

MLE for σ^2

$X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$.

Consider $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$.

$\widehat{\sigma^2}_{\text{MLE}} = \frac{1}{n} S_{XX} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is biased so we can use $\widehat{\sigma^2}_{\text{U}} = \frac{1}{n-1} S_{XX}$. Are we done?

To minimize the MSE of an estimator for σ^2, we can minimize

$$\mathbb{E}[(f(\{X_i\}_{i=1}^{n}) - \sigma^2)^2]$$

Let's try to minimize the MSE

$X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$.
Consider estimators that depend (linearly) on S_{XX}.

$$\begin{aligned}
\mathrm{MSE}[\lambda S_{XX}] &= \mathbb{E}[(\lambda S_{XX} - \sigma^2)^2] && (4)\\
&= \mathbb{E}[(\lambda S_{XX})^2 + \sigma^4 - 2\sigma^2 \lambda S_{XX}] && (5)\\
&= \lambda^2[(n-1)^2\sigma^4 + 2(n-1)\sigma^4] - 2(n-1)\sigma^4\lambda + \sigma^4. && (6)
\end{aligned}$$

Set $\lambda = \frac{1}{n+1}$ to minimize the MSE.

Statistic that minimizes the MSE

$X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$.
$\frac{S_{XX}}{n+1}$ minimizes the MSE of the variance estimator. It is:

1. Not the MLE!
2. Not unbiased!

So now what? Rao-Blackwell could help us see what is happening.

Theorem (Rao (1945), Blackwell (1947))

Let $\hat{\theta}$ be an estimator of θ where $\mathbb{E}[\hat{\theta}^2] < \infty$. Suppose that T is sufficient for θ, and let $\theta^* = \mathbb{E}[\hat{\theta} \mid T]$. Then

$$\mathbb{E}[(\theta^* - \theta)^2] \leq \mathbb{E}[(\hat{\theta} - \theta)^2].$$

(The inequality is strict unless $\hat{\theta}$ is a function of T.)

Proof of Rao–Blackwell(–Kolmogorov) Theorem

$$\begin{aligned}
\mathbb{E}[(\theta^* - \theta)^2] &= \mathbb{E}[(\mathbb{E}[\hat{\theta} \mid T] - \theta)^2] && (7)\\
&= \mathbb{E}[(\mathbb{E}[\hat{\theta} - \theta \mid T])^2] && (8)\\
&\leq \mathbb{E}[\mathbb{E}[(\hat{\theta} - \theta)^2 \mid T]] && (9)\\
&= \mathbb{E}[(\hat{\theta} - \theta)^2], && (10)
\end{aligned}$$

where we used the law of total expectation and Jensen's inequality.

Sufficient Statistic Perturbation (SSP)

1. Add *carefully calibrated* noise to sufficient statistics to satisfy differential privacy (Dwork et al. 2006).
2. For small datasets or homogeneous datasets, it might be best to use robust methods.
3. But eventually (with large enough sample size or high variance), Rao-Blackwell kicks in.

Consider simple linear regression: $Y \sim \mathcal{N}(X\beta, \sigma_e^2 I_{n\times n})$, where the design matrix is of the form $X \in \mathbb{R}^{n\times 2}$ from n observations, where

$$X = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \cdots & \cdots \\ 1 & x_{n-1} - \bar{x} \\ 1 & x_n - \bar{x} \end{pmatrix}, \quad \beta = (\beta_1, \beta_2), \quad Y \in \mathbb{R}^n,$$

and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

$$X^T X = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2. \end{pmatrix}$$

For some $\rho, \Delta > 0$, the private estimate of $X^T X$ is

$$\widetilde{X^T X} = X^T X + N, \quad N \sim \mathcal{N}\left(0, \frac{\Delta^2}{\rho^2} I_{2\times 2}\right).$$

Without privacy, $\hat{\beta} = (X^T X)^{-1} X^T Y$ is an estimate of β.

Consider simple linear regression: $Y \sim \mathcal{N}(X\beta, \sigma_e^2 I_{n\times n})$.

$$X^T X = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2. \end{pmatrix}$$

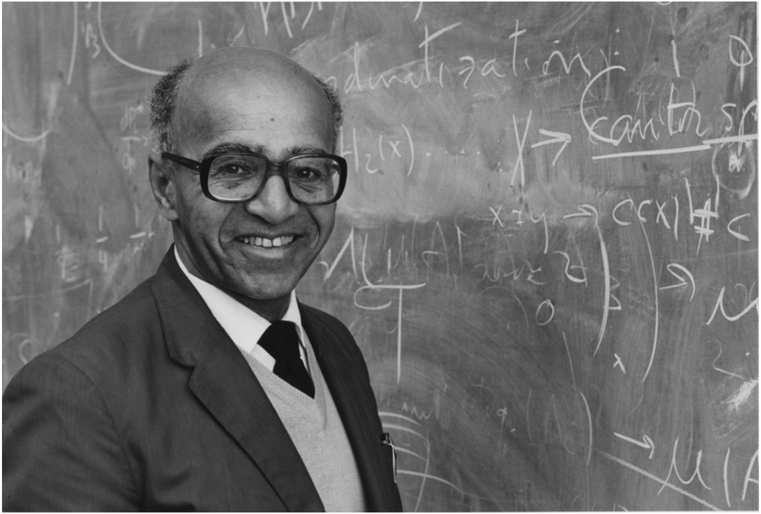
For some $\rho, \Delta > 0$, the private estimate of $X^T X$ is

$$\widetilde{X^T X} = X^T X + N, \quad N \sim \mathcal{N}\left(0, \frac{\Delta^2}{\rho^2} I_{2\times 2}\right).$$

1. The smaller n is, the more inaccurate $\widetilde{X^T X}$ will be.
2. The smaller the variance is, the more inaccurate $\widetilde{X^T X}$ will be.

In such cases, use robust methods.
But as n or the variance gets larger, Rao-Blackwell kicks in!

The World Needs More Blackwells (1919 — 2010)!



Thanks — Some References

1. *Rao* (1945):
   Information and accuracy attainable in the estimation of statistical parameters.
2. *Blackwell* (1947):
   Conditional expectation and unbiased sequential estimation.
3. *Alabi, McMillan, Sarathy, Smith, Vadhan* (2020):
   Differentially Private Simple Linear Regression.
4. *Alabi, Vadhan* (2022):
   Hypothesis Testing for Differentially Private Linear Regression.