Non-Parametric Up-and-Down Experimentation
Revisited[1]

Michael N. Katehakis

Rutgers Business School

18TH ANNUAL APPLIED PROBABILITY DAY:
IN HONOR AND MEMORY OF CYRUS DERMAN

THE CENTER FOR APPLIED PROBABILITY AT COLUMBIA UNIVERSITY
December 2, 2011

[1] Joint work with Cyrus Derman

The Basic Problem

Derman C. (1957). Non-Parametric Up-and-down Experimentation *The Annals of Mathematical Statistics* , 28(3), pp. 795-798.

- Let $Y(x)$ be a random variable such that:

$$Y(x) = \begin{cases} 1 & P(Y(x) = 1) = F(x) \\ 0 & P(Y(x) = 0) = 1 - F(x) \end{cases}$$

  where $F(x)$ is an unknown distribution function.

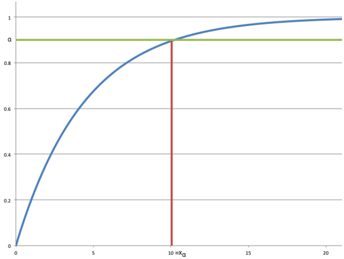
- Objective: given α estimate the α- quantile of $F(x)$,

$$x_a = F^{-1}(\alpha)$$

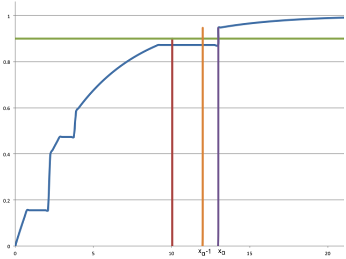
  with observations distributed like $Y(x)$ where the choice of x is under control.

- Special Case: Median: $x_{0.50} = L_{0.50}$, $\alpha = .50$.

Possible Cases



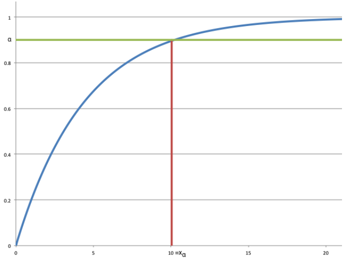
$$F(x_\alpha) = \alpha$$



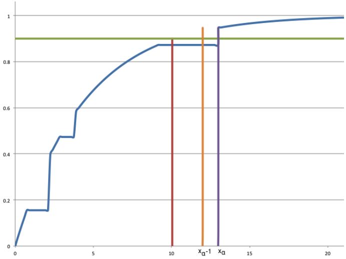
$$x_\alpha = \inf\{x \in \mathbf{N} \ : \ F(x) \geq \alpha\}$$

$$\underline{x}_\alpha = \sup\{x \in \mathbf{N} \ : \ F(x) \leq \alpha\}$$

Possible Cases



$$F(x_\alpha) = \alpha$$



$$\overline{x}_\alpha = \inf\{x \in \mathbf{N} \ : \ F(x) \geq \alpha\}$$

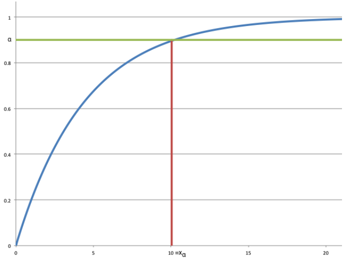
$$\underline{x}_\alpha = \sup\{x \in \mathbf{N} \ : \ F(x) \leq \alpha\}$$

Objective:

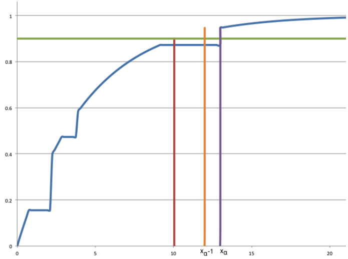
Start with x_0, observe $Y(x_0) = 1, 0$, **Pr**$(Y(x_0) = 1) = F(x_0)$.

choose X_1, observe $Y(X_1) = 1, 0$, **Pr**$(Y(X_1) = 1) = F(X_1)$.

Possible Cases



$$F(x_\alpha) = \alpha$$

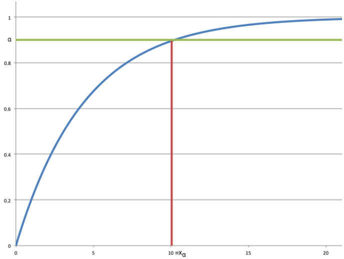


$$\bar{x}_\alpha = \inf\{x \in \mathbf{N} \ : \ F(x) \geq \alpha\}$$
$$\underline{x}_\alpha = \sup\{x \in \mathbf{N} \ : \ F(x) \leq \alpha\}$$

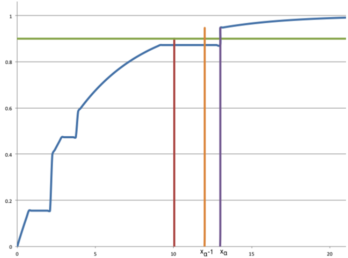
Objective:

Start with	x_0,	observe $Y(x_0) = 1, 0$,	**Pr**$(Y(x_0) = 1) = F(x_0)$.
choose	X_1,	observe $Y(X_1) = 1, 0$,	**Pr**$(Y(X_1) = 1) = F(X_1)$.
	$\ldots$,	$\ldots$	

Possible Cases



$$F(x_\alpha) = \alpha$$



$$x_\alpha = \inf\{x \in \mathbf{N} \;:\; F(x) \geq \alpha\}$$

$$\underline{x}_\alpha = \sup\{x \in \mathbf{N} \;:\; F(x) \leq \alpha\}$$

Objective:

Start with	x_0,	observe $Y(x_0) = 1, 0$,	**Pr**$(Y(x_0) = 1) = F(x_0)$.
choose	X_1,	observe $Y(X_1) = 1, 0$,	**Pr**$(Y(X_1) = 1) = F(X_1)$.
	$\ldots,$	$\ldots$	
choose	X_n,	such that $\hat{x}_\alpha(x_0, X_1, \ldots, X_n) \xrightarrow{a.s.} x_\alpha$.	

Newsvendor Model

- When x items are ordered in a period we observe if there is : no shortage $Y(x) = 1$ or shortage $Y(x) = 0$.
- The optimal order quantity x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha \ = (p - c)/p\}$
- F is the demand distribution.

Relevance

Newsvendor Model

- When x items are ordered in a period we observe if there is : no shortage $Y(x) = 1$ or shortage $Y(x) = 0$.
- The optimal order quantity x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha \ = (p - c)/p\}$
- F is the demand distribution.

Drug Testing

- Dosage is x
- Observe a success $Y(x) = 1$ or failure $Y(x) = 0$.
- The optimal dosage x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha\}$
- F is the quantal response function.

Relevance

Newsvendor Model

- When x items are ordered in a period we observe if there is : no shortage $Y(x) = 1$ or shortage $Y(x) = 0$.
- The optimal order quantity x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha \ = (p-c)/p\}$
- F is the demand distribution.

Drug Testing

- Dosage is x
- Observe a success $Y(x) = 1$ or failure $Y(x) = 0$.
- The optimal dosage x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha\}$
- F is the quantal response function.

Educational Testing

- Difficulty level of a test question is x
- Observe a correct $Y(x) = 1$ or wrong $Y(x) = 0$ answer.
- The student's ability is the x_α for which $F(x_\alpha) = \alpha$

Relevance

Newsvendor Model

- When x items are ordered in a period we observe if there is : no shortage $Y(x) = 1$ or shortage $Y(x) = 0$.
- The optimal order quantity x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha \ = (p-c)/p\}$
- F is the demand distribution.

Drug Testing

- Dosage is x
- Observe a success $Y(x) = 1$ or failure $Y(x) = 0$.
- The optimal dosage x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha\}$
- F is the quantal response function.

Educational Testing

- Difficulty level of a test question is x
- Observe a correct $Y(x) = 1$ or wrong $Y(x) = 0$ answer.
- The student's ability is the x_α for which $F(x_\alpha) = \alpha$

Manufacturing

Relevance

Newsvendor Model

- When x items are ordered in a period we observe if there is : no shortage $Y(x) = 1$ or shortage $Y(x) = 0$.
- The optimal order quantity x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha \ = (p-c)/p\}$
- F is the demand distribution.

Drug Testing

- Dosage is x
- Observe a success $Y(x) = 1$ or failure $Y(x) = 0$.
- The optimal dosage x_α is determined by a quantile requirement: $x_\alpha = \inf\{x \in \mathbb{N} \ : \ F(x) \geq \alpha\}$
- F is the quantal response function.

Educational Testing

- Difficulty level of a test question is x
- Observe a correct $Y(x) = 1$ or wrong $Y(x) = 0$ answer.
- The student's ability is the x_α for which $F(x_\alpha) = \alpha$

Manufacturing

The Sensitivity of an Explosive

Derman's Up & Down Method

- Grid or experimental range of x to a set of numbers of the form

$$b + hn \;\; (-\infty < b < \infty,\; h > 0,\; n = 0, \pm 1, \ldots).$$

  For convenience one can assume $b = 0$, $h = 1$.

- Procedure:
  Start with x_0 (init. guess) $y_0 = Y(x_0)$ is observed
  where $P(Y(x_0) = 1) = F(x_0) = 1 - P(Y(x_0) = 0)$.
  Given x_0, $y(x_0)$, $\ldots$, x_{n-1}, $y(x_{n-1})$ for $n \geq 0$, define:

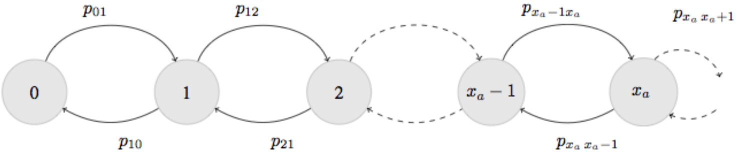
$$x_{n+1} = \begin{cases} x_n + 1 \text{ if } & \begin{cases} y(x_n) = 0 & \text{with probability: } 1 \\ y(x_n) = 1 & \text{with probability: } 1 - \frac{1}{2\alpha} \end{cases} \\ x_n - 1 \text{ if } & \quad y(x_n) = 1 \qquad \text{with probability: } \frac{1}{2\alpha} \end{cases}$$

  where w.l.o.g. $\alpha > 1/2$

- The estimate $\hat{x}_{\alpha,n}$ of x_α based on n observations is

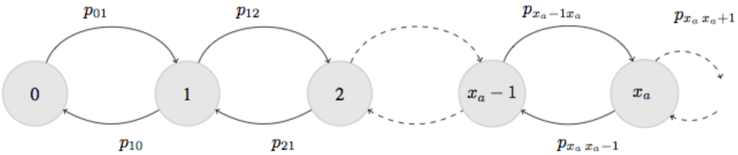
$$\hat{x}_{\alpha,n} = \begin{cases} \text{the } \textit{most frequent value} \text{ of } x\text{'}s, & \text{if unique,} \\ \text{the arithmetic average of the most frequent levels} & \text{otherwise.} \end{cases}$$

Derman's Main Result



$\{X_n,\}_{n\geq 1}$ process: $p_{i,i+1} = 1 - F(i)/2\alpha$

Derman's Main Result

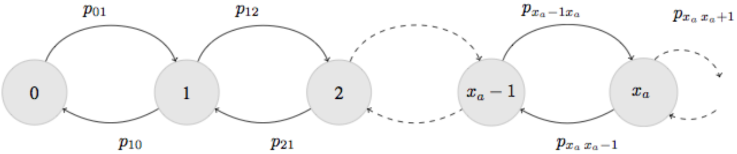


$\{X_n,\}_{n\geq 1}$ process: $p_{i,i+1} = 1 - F(i)/2\alpha$

Condition A: If $F(x)$ is strictly increasing for $x \in [x_\alpha - 1, x_\alpha]$ then:

$$\mathbf{Pr}(max(|\overline{\lim_n}\,\hat{x}_{\alpha,n} - x_\alpha|, |\underline{\lim_n}\,\hat{x}_{\alpha,n} - x_\alpha|) < 1) = 1$$

Derman's Main Result



$\{X_n,\}_{n\geq 1}$ process: $p_{i,i+1} = 1 - F(i)/2\alpha$

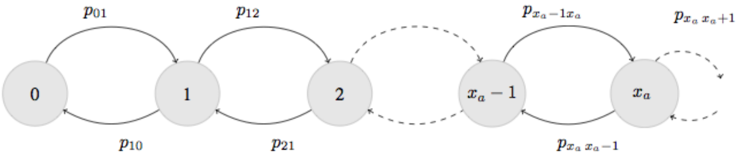
Condition A: If $F(x)$ is strictly increasing for $x \in [x_\alpha - 1, x_\alpha]$ then:

$$\textbf{Pr}(max(|\overline{\lim_n}\,\hat{x}_{\alpha,n} - x_\alpha|, |\underline{\lim_n}\,\hat{x}_{\alpha,n} - x_\alpha|) < 1) = 1$$

Main Tool: $\pi(x) = \lim_n \textbf{Pr}(X_n = x\,|\,x_o)$

$$\pi(0) \leq \pi(1) \leq \ldots \leq \pi(x_\alpha - 1) < \pi(x_\alpha) \geq \pi(x_\alpha + 1) \geq \pi(x_\alpha + 2) \geq \ldots$$

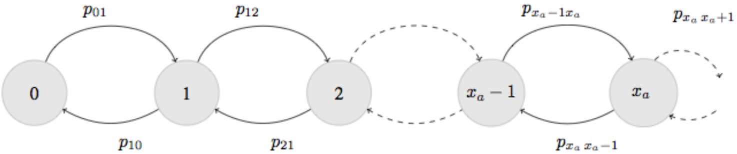
Derman's U-D Revisited



$\{X_n,\}_{n\geq 1}$ process: $p_{i,i+1} = 1 - F(i)/2\alpha$

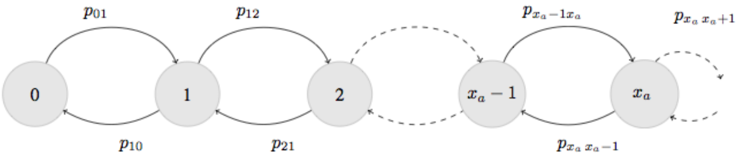
- What is an efficient *Stopping Criterion ?*
- What is the *Error Probability ?*
- *What if Condition A does not hold ?*

Derman's U-D Revisited - Answers



$\{X_n,\}_{n\geq 1}$ process: $p_{i,i+1} = 1 - F(i)/2\alpha$

Derman's U-D Revisited - Answers

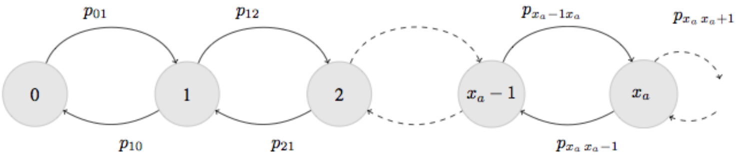


$\{X_n,\}_{n\geq 1}$ process: $p_{i,i+1} = 1 - F(i)/2\alpha$

Observation 1 if x_α is on the grid

$p_{00} = 1 \geq p_{12} \geq \ldots \geq p_{x_\alpha - 1 x_\alpha} \geq p_{x_\alpha x_\alpha + 1} = 1/2 \geq p_{x_\alpha + 1 x_\alpha + 2} \geq \ldots$

Derman's U-D Revisited - Answers



$\{X_n,\}_{n\geq 1}$ process: $p_{i,i+1} = 1 - F(i)/2\alpha$

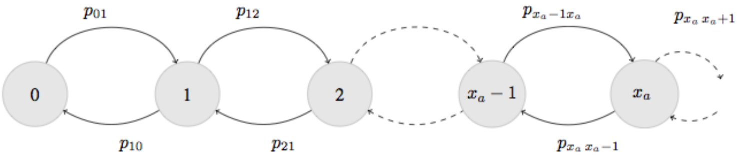
Observation 1 if x_α is on the grid

$$p_{00} = 1 \geq p_{12} \geq \ldots \geq p_{x_\alpha - 1 x_\alpha} \geq p_{x_\alpha x_\alpha + 1} = 1/2 \geq p_{x_\alpha + 1 x_\alpha + 2} \geq \ldots$$

Observation 2 if x_α is not on the grid:

$$p_{00} = 1 \geq p_{12} \geq \ldots \geq p_{x_\alpha - 1 x_\alpha} > 1/2 \geq p_{x_\alpha x_\alpha + 1} \geq p_{x_\alpha + 1 x_\alpha + 2} \geq \ldots$$

Derman's U-D Revisited - Answers



$\{X_n,\}_{n\geq 1}$ process: $p_{i,i+1} = 1 - F(i)/2\alpha$

Observation 1 if x_α is on the grid

$$p_{00} = 1 \geq p_{12} \geq \ldots \geq p_{x_\alpha - 1 x_\alpha} \geq p_{x_\alpha x_\alpha + 1} = 1/2 \geq p_{x_\alpha + 1 x_\alpha + 2} \geq \ldots$$

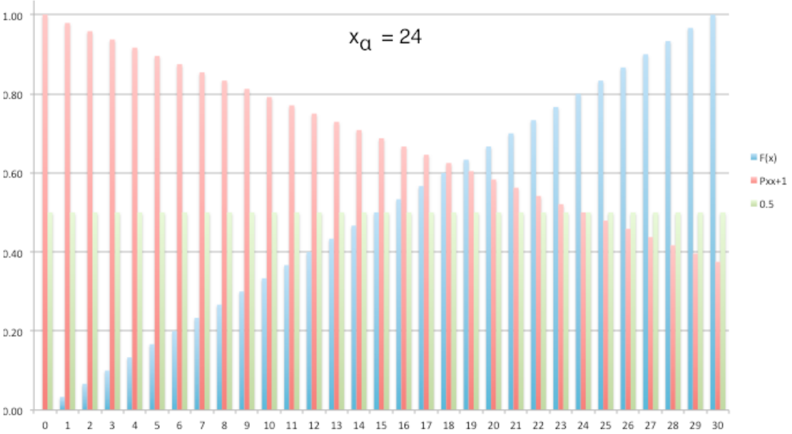
Observation 2 if x_α is not on the grid:

$$p_{00} = 1 \geq p_{12} \geq \ldots \geq p_{x_\alpha - 1 x_\alpha} > 1/2 \geq p_{x_\alpha x_\alpha + 1} \geq p_{x_\alpha + 1 x_\alpha + 2} \geq \ldots$$

Compare with :

$$\pi(0) \leq \pi(1) \leq \ldots \leq \pi(x_\alpha - 1) \leq \pi(x_\alpha) \geq \pi(x_\alpha + 1) \geq \pi(x_\alpha + 2) \geq \ldots$$

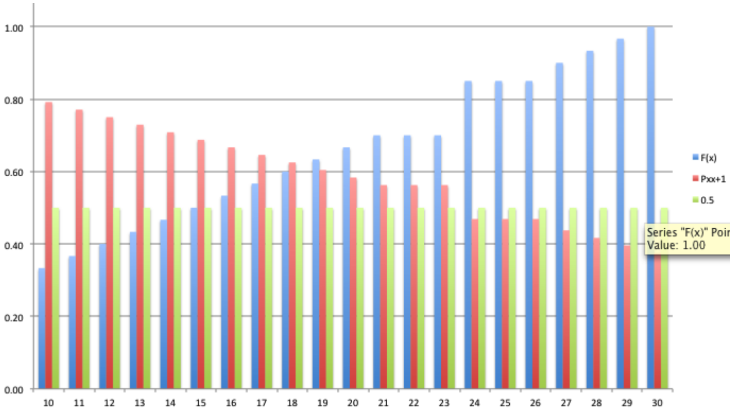
Derman's U-D Revisited -Continued



$\alpha = .0.80$, $x_\alpha = 24$ **on the grid:** $F(24) = .8$

$p_{23,24} > 1/2 = p_{24,25} > p_{25,26} < 1/2$

Derman's U-D Revisited -Continued



$\alpha = .0.80$, $x_\alpha = 24$ **not on the grid:** $F(23) < 0.8 < F(24)$

$$p_{23,24} > 1/2 > p_{24,25} > p_{25,26} > \dots$$

Derman and Katehakis (2010-2011) New Results

For simplicity consider the case x_α is on the grid.

- We can **Stop** the procedure:
  - $\tau^1_\alpha = \inf\{k \ : \ \hat{p}_{k,k+1}(x_0, \ldots, X_k) \in (1/2 - \epsilon, 1/2 + \epsilon)\}$
  - $\tau^2_\alpha = \inf\{k \ : \ \hat{p}_{k,k+1}(x_0, \ldots, X_k) \in (1/2 - \epsilon, 1/2 + \epsilon) \ \& \ V_k = \max_{k'}\{V_{k'}\} \ \}$
  - $\hat{x}_a = X_{\tau^i_\alpha}$
- Have $\mathbf{Pr}(\tau^i_\alpha > u) = c^i_F(e^{-u} + \epsilon^i_F(n, u))$ where $\epsilon^i_F(n, u)) \to 0$ $(n \to \infty)$
- We can modify the procedure by taking a second sample at τ_α (Two Stage)
- Working on using techniques of Adaptive MDPs to obtain:

$$R^\pi_N \geq R^{\pi^{DK}}_N = M_{DK}(P)logN + o(logN)$$

Derman and Katehakis (2010-2011) New Results

For simplicity consider the case x_α is on the grid.

- We can **Stop** the procedure:
  - $\tau_\alpha^1 = \inf\{k \ : \ \hat{p}_{k,k+1}(x_0,\ldots,X_k) \in (1/2-\epsilon, 1/2+\epsilon)\}$
  - $\tau_\alpha^2 = \inf\{k \ : \ \hat{p}_{k,k+1}(x_0,\ldots,X_k) \in (1/2-\epsilon, 1/2+\epsilon)$ & $V_k = \max_{k'}\{V_{k'}\} \ \}$
  - $\hat{x}_a = X_{\tau_\alpha^i}$
- Have $\mathbf{Pr}(\tau_\alpha^i > u) = c_F^i(e^{-u} + \epsilon_F^i(n,u))$ where $\epsilon_F^i(n,u)) \to 0$ $(n \to \infty)$
- We can modify the procedure by taking a second sample at τ_α (Two Stage)
- Working on using techniques of Adaptive MDPs to obtain:

$$R_N^\pi \geq R_N^{\pi^{DK}} = M_{DK}(P)logN + o(logN)$$

Similar results hold in the case x_α is not on the grid or Condition A does not hold.

Background

Non Parametric "Up & Down" Methods

- Anderson, T., McCarthy, P., Tukey, J., (1946). Staircase method of sensitivity testing. *Naval Ordinance Report 65-46*, Statistical Research Group, Princeton University, Princeton, NJ.
- Dixon, W.J. and Mood, A.M. (1948). A method for obtaining and analyzing sensitive data. *Journal of the American Statistical Association* 43, pp. 109 -126. *The validity of their procedure depends on the assumption that $F(x)$ is normal.*
- Derman C. (1957). Non-Parametric Up-and-down Experimentation
- Wetherill, G.B. (1963). Sequential estimation of quantal response curves.

Non Parametric "Up & Down" Methods

- Anderson, T., McCarthy, P., Tukey, J., (1946). Staircase method of sensitivity testing. *Naval Ordinance Report 65-46*, Statistical Research Group, Princeton University, Princeton, NJ.
- Dixon, W.J. and Mood, A.M. (1948). A method for obtaining and analyzing sensitive data. *Journal of the American Statistical Association* 43, pp. 109 -126. *The validity of their procedure depends on the assumption that $F(x)$ is normal.*
- Derman C. (1957). Non-Parametric Up-and-down Experimentation
- Wetherill, G.B. (1963). Sequential estimation of quantal response curves.
- Wetherhill, G.B., Glazebrook, K.D., (1986). *Sequential Methods in Statistics*, Chapman & Hall, London.
- Durham, D.S., Flournoy N. and Rosenberger W. F (1997). A Random Walk Rule for Phase I Clinical Trials
- Bortot P. and Giovagnolia A. (2005). Up-and-down experiments of first and second order
- Ivanova A(2006). Dose-Finding in Oncology-Nonparametric Methods,
- Pollak, R., Palazotto, A., Nicholas, T. (2006). A simulation-based investigation of the stair- case method for fatigue strength testing.
- Baldi Antognini A. and . Giovagnoli A (2010). Compound Optimal Allocation for Individual and Collective Ethics in Binary Clinical Trials,

Background: "Stochastic Approximation" and Related Methods

- Robbins, H. and Monro, S. (1951). A stochastic approximation method, *The Annals of Mathematical Statistics*, 22(3), pp. 400–407.
- Derman C. and J. Sacks (1959). On Dvoretzky's Stochastic Approximation Theorem, *The Annals of Mathematical Statistics*, 30(2), pp. 601-606.
- Frederic M. Lord (1971) Tailored Testing, An Application of Stochastic Approximation *Journal of the American Statistical Association* Vol. 66, No. 336, pp. 707-711.

The Robbins and Monro SA scheme can be used for estimating any quantile and it imposes no parametric assumptions on $F(x)$.

- The method does assume, however, that the range of possible experimental values of x is the real line.
- It has slow convergence rate.

Background: Optimal Adaptive Policies for MABs

- Gittins, J.C. and Jones, D.M. (1979). "A dynamic allocation index for the discounted multiarmed bandit problem",
- Katehakis M. N. and A. F. Veinott Jr. (1987). "The Multi-Armed Bandit problem: decomposition and computation",

Background: Optimal Adaptive Policies for MABs

- Gittins, J.C. and Jones, D.M. (1979). "A dynamic allocation index for the discounted multiarmed bandit problem",
- Katehakis M. N. and A. F. Veinott Jr. (1987). "The Multi-Armed Bandit problem: decomposition and computation",
- Lai T.L. and H. E. Robbins (1985) "Asymptotically Efficient Adaptive Allocation Rules",

$$R_N^\pi = \max\{\mu_1, \ldots, \mu_n\}N - V_N^\pi$$

$$R_N^\pi \geq R_N^{\pi^{LR}} = M_{LR}(P)logN + o(logN)$$

- Note:

$$R_N^{\pi^0} = M_{LR}(P)logN + o(logN) = o(N^a) \quad \forall a > 0$$

- Even better:

$$\overline{\lim_N} R_N^{\pi^0} / R_N^\pi \leq 1$$

- Katehakis M. N. and H. E. Robbins (1995). "Sequential choice from several populations",

Background: Optimal Adaptive Policies for MDPs

- Burnetas, A.N. and M. N. Katehakis (1996) "Optimal Adaptive Policies for Sequential Allocation Problems",
- Burnetas, A.N. and M. N. Katehakis (1997) "Optimal Adaptive Policies for Markov Decision Processes",

$$R_N^{\pi} = V_N - V_N^{\pi}$$

$$R_N^{\pi} \geq {R_N^{\pi}}^{BK} = M_{BK}(P) logN + o(logN)$$

Background: Optimal Adaptive Policies for MDPs

- Burnetas, A.N. and M. N. Katehakis (1996) "Optimal Adaptive Policies for Sequential Allocation Problems",
- Burnetas, A.N. and M. N. Katehakis (1997) "Optimal Adaptive Policies for Markov Decision Processes",

$$R_N^{\pi} = V_N - V_N^{\pi}$$

$$R_N^{\pi} \geq R_N^{\pi^{BK}} = M_{BK}(P) logN + o(logN)$$

Near-optimal Regret Bounds for Reinforcement Learning

- Auer, P. and R. Ortner (2007) Logarithmic online regret bounds for undiscounted reinforcement learning.

$$R_N^{\pi} \geq R_N^{\pi^{AO}} = M_{AO}(P) logN + o(logN)$$

- Tewari A. and P. Bartlett (2008). "Optimistic linear programming gives logarithmic regret for irreducible MDPs",

$$R_N^{\pi} \geq R_N^{\pi^{TB}} = M_{TB}(P) logN + o(logN)$$

- Auer, P. and Jaksch, T. and Ortner, R. (2009). "Near-optimal regret bounds for reinforcement learning",

Background: Optimal Adaptive Policies for MDPs

- Burnetas, A.N. and M. N. Katehakis (1996) "Optimal Adaptive Policies for Sequential Allocation Problems",
- Burnetas, A.N. and M. N. Katehakis (1997) "Optimal Adaptive Policies for Markov Decision Processes",

$$R_N^\pi = V_N - V_N^\pi$$

$$R_N^\pi \geq {R_N^\pi}^{BK} = M_{BK}(P) logN + o(logN)$$

Near-optimal Regret Bounds for Reinforcement Learning

- Auer, P. and R. Ortner (2007) Logarithmic online regret bounds for undiscounted reinforcement learning.

$$R_N^\pi \geq {R_N^\pi}^{AO} = M_{AO}(P) logN + o(logN)$$

- Tewari A. and P. Bartlett (2008). "Optimistic linear programming gives logarithmic regret for irreducible MDPs",

$$R_N^\pi \geq {R_N^\pi}^{TB} = M_{TB}(P) logN + o(logN)$$

- Auer, P. and Jaksch, T. and Ortner, R. (2009). "Near-optimal regret bounds for reinforcement learning",

$$M_{AO}(P) > M_{TB}(P) > M_{BK}(P) \; \forall P$$

Background: Optimal Adaptive Policies for MDPs

- Burnetas, A.N. and M. N. Katehakis (1996) "Optimal Adaptive Policies for Sequential Allocation Problems",
- Burnetas, A.N. and M. N. Katehakis (1997) "Optimal Adaptive Policies for Markov Decision Processes",

$$R_N^\pi = V_N - V_N^\pi$$

$$R_N^\pi \geq {R_N^\pi}^{BK} = M_{BK}(P) logN + o(logN)$$

Near-optimal Regret Bounds for Reinforcement Learning

- Auer, P. and R. Ortner (2007) Logarithmic online regret bounds for undiscounted reinforcement learning.

$$R_N^\pi \geq {R_N^\pi}^{AO} = M_{AO}(P) logN + o(logN)$$

- Tewari A. and P. Bartlett (2008). "Optimistic linear programming gives logarithmic regret for irreducible MDPs",

$$R_N^\pi \geq {R_N^\pi}^{TB} = M_{TB}(P) logN + o(logN)$$

- Auer, P. and Jaksch, T. and Ortner, R. (2009). "Near-optimal regret bounds for reinforcement learning",

$$M_{AO}(P) > M_{TB}(P) > M_{BK}(P) \; \forall P$$

Background: Optimal Adaptive Policies for MDPs

- Burnetas, A.N. and M. N. Katehakis (1996) “Optimal Adaptive Policies for Sequential Allocation Problems”,
- Burnetas, A.N. and M. N. Katehakis (1997) ”Optimal Adaptive Policies for Markov Decision Processes”,

$$R_N^\pi = V_N - V_N^\pi$$

$$R_N^\pi \geq R_N^{\pi^{BK}} = M_{BK}(P) logN + o(logN)$$

Near-optimal Regret Bounds for Reinforcement Learning

- Auer, P. and R. Ortner (2007) Logarithmic online regret bounds for undiscounted reinforcement learning.

$$R_N^\pi \geq R_N^{\pi^{AO}} = M_{AO}(P) logN + o(logN)$$

- Tewari A. and P. Bartlett (2008). “Optimistic linear programming gives logarithmic regret for irreducible MDPs”,

$$R_N^\pi \geq R_N^{\pi^{TB}} = M_{TB}(P) logN + o(logN)$$

- Auer, P. and Jaksch, T. and Ortner, R. (2009). “Near-optimal regret bounds for reinforcement learning”,

$$M_{AO}(P) > M_{TB}(P) > M_{BK}(P) \ \forall P$$

The Robbins Monro Method for x_α

Compute x_α such that

$$M(x_\alpha) = \alpha$$

- Deterministic case $M(x)$ is **known** solution by:

$$x_n = x_{n-1} + a_n[\alpha - M(x_{n-1})],$$

- Stochastic case $M(x)$ is **unknown**

$$E(Y(x)) = M(x) = \alpha. \ \forall \ x$$

  Solution by:

$$x_n = x_{n-1} + a_n[\alpha - y_{n-1})],$$

  The estimate $\hat{x}_{\alpha,n}$ of x_α based on n observations is $\hat{x}_{\alpha,n} = x_n$,
  Under regularity conditions[2]

[2] E.g., M is non-decreasing, there exists a solution $M(x_\alpha) = \alpha$, $\exists \frac{M(x)}{dx} > 0$ at x_α, and $\sum_{n=0}^{\infty} a_n = \infty$, $\sum_{n=0}^{\infty} a_n^2 < \infty$

The Dixon and Mood Method for LD_{50} ($x_{0.50}$)

- Grid or experimental range[3] of x to a set of numbers of the form

$$b + hn \quad (-\infty < b < \infty,\ h > 0,\ n = 0, \pm 1, \ldots).$$

  For convenience one can assume $b = 0,\ h = 1$.

- Data: Treatments are administered sequentially at dosage: X_i, as follows:
  Start with x_0 (arbitrary guess) $y_0 = Y(x_0)$ is observed where
  $P(Y(x_0) = 1) = F(x_0) = 1 - P(Y(x_0) = 0)$.
  Given $x_0,\ y(x_0),\ \ldots,\ x_k,\ y(x_{n-1})$ for $n \geq 0$, define recursively

$$x_n = \begin{cases} x_{n-1} + 1 \text{ if } & y(x_k) = 0, \\ x_{n-1} - 1 \text{ if } & y(x_k) = 1. \end{cases}$$

- The estimate $\hat{x}_\alpha$ of x_α based on n observations is

$$\hat{x}_{0.50} = \frac{1}{n+1} \sum_{j=1}^{n+1} x_j,$$

[3] Natural limitations such as when x is obtained by a counting procedure - limitations on the precision of measuring instruments.