Markov decision process under mean-field interaction and application to targeted advertising

Médéric MOTTE

Université de Paris, LPSM

Based on joint work with
Huyên PHAM, Université de Paris, LPSM

European summer school in financial mathematics
Padova, Sept 2-6, 2019

Mean-field approach to large population stochastic control

- Large number of N interacting dynamic agents

Mean-field approach to large population stochastic control

- Large number of N interacting dynamic agents
- Center of decision on top of the population

Mean-field approach to large population stochastic control

- Large number of N interacting dynamic agents
- Center of decision on top of the population
  - **Social planner** aims to optimize the global gain/cost of collectivity
    $\rightarrow$ Pareto efficiency $\neq$ Nash equilibrium

Mean-field approach to large population stochastic control

- Large number of N interacting dynamic agents
- Center of decision on top of the population
  - **Social planner** aims to optimize the global gain/cost of collectivity $\rightarrow$ Pareto efficiency $\neq$ Nash equilibrium
  - **Influencer** control the state of all players, e.g. through advertising in models of herd behavior/connected people

Mean-field approach to large population stochastic control

- Large number of N interacting dynamic agents
- Center of decision on top of the population
  - **Social planner** aims to optimize the global gain/cost of collectivity $\rightarrow$ Pareto efficiency $\neq$ Nash equilibrium
  - **Influencer** control the state of all players, e.g. through advertising in models of herd behavior/connected people

▶ Our main focus:

Mean-field approach to large population stochastic control

- Large number of N interacting dynamic agents
- Center of decision on top of the population
  - **Social planner** aims to optimize the global gain/cost of collectivity $\rightarrow$ Pareto efficiency $\neq$ Nash equilibrium
  - **Influencer** control the state of all players, e.g. through advertising in models of herd behavior/connected people

▶ Our main focus:

- **Discrete time** and space

Mean-field approach to large population stochastic control

- Large number of N interacting dynamic agents
- Center of decision on top of the population
  - **Social planner** aims to optimize the global gain/cost of collectivity
    $\rightarrow$ Pareto efficiency $\neq$ Nash equilibrium
  - **Influencer** control the state of all players, e.g. through advertising in models of herd behavior/connected people

- Our main focus:
  - **Discrete time** and space
  - When $N \rightarrow \infty$: McKean-Vlasov Markov Decision Process (MKV-MDP)

A motivating targeted advertising example

- A (phone) company C (the influencer),

A motivating targeted advertising example

- A (phone) company C (the influencer),
- A social network SN

A motivating targeted advertising example

- A (phone) company C (the influencer),
- A social network SN
  - N connected users of SN: state $=$ client or not of C

A motivating targeted advertising example

- A (phone) company C (the influencer),
- A social network SN
  - N connected users of SN: state $=$ client or not of C
  - Users data: cookies

A motivating targeted advertising example

- A (phone) company C (the influencer),
- A social network SN
  - N connected users of SN: state $=$ client or not of C
  - Users data: cookies

▶ **Targeted advertising**: SN can display ads for C to some users according to their cookies.

What is the (best) targeted strategy?

- For each user i, cookies:

What is the (best) targeted strategy?

- For each user i, cookies:
  - Initial information Γ^i

What is the (best) targeted strategy?

- For each user i, cookies:
  - Initial information Γ^i
  - At any time t, additional information ε^i_t, e.g. time spent on a forum discussing about phones

What is the (best) targeted strategy?

- For each user i, cookies:
  - Initial information Γ^i
  - At any time t, additional information ε_t^i, e.g. time spent on a forum discussing about phones

- A program is implemented on each laptop based on the cookies for displaying ads. This program is the same on each laptop

What is the (best) targeted strategy?

- For each user i, cookies:
  - Initial information Γ^i
  - At any time t, additional information ε_t^i, e.g. time spent on a forum discussing about phones

- A program is implemented on each laptop based on the cookies for displaying ads. This program is the same on each laptop

$\rightarrow$ Same policy: $\alpha_t^i = \pi_t(\Gamma^i, \varepsilon_1^i, \ldots, \varepsilon_t^i)$

What is the (best) targeted strategy?

- For each user i, cookies:
  - Initial information Γ^i
  - At any time t, additional information ε_t^i, e.g. time spent on a forum discussing about phones

- A program is implemented on each laptop based on the cookies for displaying ads. This program is the same on each laptop
$\rightarrow$ Same policy: $\alpha_t^i = \pi_t(\Gamma^i, \varepsilon_1^i, \ldots, \varepsilon_t^i)$

Remark: The program does not have access to the states of the individuals, but to the cookies: $\rightarrow$ "Open-loop" policy.

What is the (best) targeted strategy?

- For each user i, cookies:
  - Initial information Γ^i
  - At any time t, additional information ε^i_t, e.g. time spent on a forum discussing about phones

• A program is implemented on each laptop based on the cookies for displaying ads. This program is the same on each laptop
$\rightarrow$ Same policy: $\alpha^i_t = \pi_t(\Gamma^i, \varepsilon^i_1, \ldots, \varepsilon^i_t)$

Remark: The program does not have access to the states of the individuals, but to the cookies: $\rightarrow$ "Open-loop" policy.

▶ Goal of the company: find the best ad-policy to be displayed by SN in order to attract the largest possible clients given ads costs.

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$
- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$

- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$

- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A

- (Discrete time) **transition dynamics**

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$
- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A
- (Discrete time) **transition dynamics**
  - Idiosyncratic noises: $(\varepsilon_t^i)_{t\in\mathbb{N}}$, for agent $i \in \mathbb{N}^*$, i.i.d. valued in W

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$

- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A

- (Discrete time) **transition dynamics**
  - Idiosyncratic noises: $(\varepsilon^i_t)_{t\in\mathbb{N}}$, for agent $i \in \mathbb{N}^*$, i.i.d. valued in W
  - Common noise: $(\varepsilon^0_t)_{t\in\mathbb{N}}$ for all agents, i.i.d. valued in W^0

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$
- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A
- (Discrete time) **transition dynamics**
  - Idiosyncratic noises: $(\varepsilon_t^i)_{t\in\mathbb{N}}$, for agent $i \in \mathbb{N}^*$, i.i.d. valued in W
  - Common noise: $(\varepsilon_t^0)_{t\in\mathbb{N}}$ for all agents, i.i.d. valued in W^0
  - F: meas. function from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(A) \times W \times W^0$ into $\mathcal{X}$

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$
- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A
- (Discrete time) **transition dynamics**
  - Idiosyncratic noises: $(\varepsilon_t^i)_{t\in\mathbb{N}}$, for agent $i \in \mathbb{N}^*$, i.i.d. valued in W
  - Common noise: $(\varepsilon_t^0)_{t\in\mathbb{N}}$ for all agents, i.i.d. valued in W^0
  - F: meas. function from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(A) \times W \times W^0$ into $\mathcal{X}$
- **Reward** on infinite horizon:

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$
- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A
- (Discrete time) **transition dynamics**
  - Idiosyncratic noises: $(\varepsilon_t^i)_{t\in\mathbb{N}}$, for agent $i \in \mathbb{N}^*$, i.i.d. valued in W
  - Common noise: $(\varepsilon_t^0)_{t\in\mathbb{N}}$ for all agents, i.i.d. valued in W^0
  - F: meas. function from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(A) \times W \times W^0$ into $\mathcal{X}$
- **Reward** on infinite horizon:
  - discount factor $\beta \in [0, 1)$

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$
- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A
- (Discrete time) **transition dynamics**
  - Idiosyncratic noises: $(\varepsilon_t^i)_{t\in\mathbb{N}}$, for agent $i \in \mathbb{N}^*$, i.i.d. valued in W
  - Common noise: $(\varepsilon_t^0)_{t\in\mathbb{N}}$ for all agents, i.i.d. valued in W^0
  - F: meas. function from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(A) \times W \times W^0$ into $\mathcal{X}$
- **Reward** on infinite horizon:
  - discount factor $\beta \in [0, 1)$
  - f: measurable function from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(A)$ into $\mathbb{R}$

Framework and notations

- A universal probability space $(\Omega, \mathcal{F}, \mathbb{P})$
- **State and control/action** (compact Polish) spaces: $\mathcal{X}$ and A
$\mathcal{P}(\mathcal{X})$, resp. $\mathcal{P}(A)$: set of probability measures on $\mathcal{X}$, resp. A
- (Discrete time) **transition dynamics**
  - Idiosyncratic noises: $(\varepsilon_t^i)_{t\in\mathbb{N}}$, for agent $i \in \mathbb{N}^*$, i.i.d. valued in W
  - Common noise: $(\varepsilon_t^0)_{t\in\mathbb{N}}$ for all agents, i.i.d. valued in W^0
  - F: meas. function from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(A) \times W \times W^0$ into $\mathcal{X}$
- **Reward** on infinite horizon:
  - discount factor $\beta \in [0, 1)$
  - f: measurable function from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(A)$ into $\mathbb{R}$

Remark: F and f do not depend on i: indistinguishable agents

Information and (randomized) open-loop policies/control

- **Information**:
  - Γ^i, $i \in \mathbb{N}^*$, i.i.d. valued in I: initial information (agent i)

Information and (randomized) open-loop policies/control

- **Information**:
  - Γ^i, $i \in \mathbb{N}^*$, i.i.d. valued in I: initial information (agent i)
  - Initial state $= \xi(\Gamma^i)$

Information and (randomized) open-loop policies/control

- **Information**:
  - Γ^i, $i \in \mathbb{N}^*$, i.i.d. valued in I: initial information (agent i)
  - Initial state $= \xi(\Gamma^i)$
- **Open-loop Policy**: measurable sequence $\pi \in \Pi^r_{OL}$ of functions π_t, $t \in \mathbb{N}$, from $I \times W^t \times (W^0)^t$ into A

Information and (randomized) open-loop policies/control

- **Information**:
  - Γ^i, $i \in \mathbb{N}^*$, i.i.d. valued in I: initial information (agent i)
  - Initial state $= \xi(\Gamma^i)$

- **Open-loop Policy**: measurable sequence $\pi \in \Pi^r_{OL}$ of functions π_t, $t \in \mathbb{N}$, from $I \times W^t \times (W^0)^t$ into $A \leftrightarrow$ Open-loop (nonanticipative) control for agent i:

$$\alpha_t^{\pi,i} \quad = \quad \pi_t(\Gamma^i, \varepsilon_1^i, \ldots, \varepsilon_t^i, \varepsilon_1^0, \ldots, \varepsilon_t^0),$$

Information and (randomized) open-loop policies/control

- **Information**:
  - Γ^i, $i \in \mathbb{N}^*$, i.i.d. valued in I: initial information (agent i)
  - Initial state $= \xi(\Gamma^i)$

- **Open-loop Policy**: measurable sequence $\pi \in \Pi^r_{OL}$ of functions π_t, $t \in \mathbb{N}$, from $I \times W^t \times (W^0)^t$ into A $\leftrightarrow$ Open-loop (nonanticipative) control for agent i:

$$\alpha_t^{\pi,i} \quad = \quad \pi_t(\Gamma^i, \varepsilon_1^i, \ldots, \varepsilon_t^i, \varepsilon_1^0, \ldots, \varepsilon_t^0),$$

Remark: each agent has only access to her own initial information, idiosyncratic and common noise, but follows the same policy.

Mean-field control problem in the N-population model

- **Mean-field controlled dynamics**:
  - Same initial state function $\xi \in L(I; \mathcal{X})$, and same open-loop policy π

Mean-field control problem in the N-population model

- **Mean-field controlled dynamics**:
  - Same initial state function $\xi \in L(I; \mathcal{X})$, and same open-loop policy π
  - State process $X^{i,N}$ of agent i

Mean-field control problem in the N-population model

- **Mean-field controlled dynamics**:
  - Same initial state function $\xi \in L(I; \mathcal{X})$, and same open-loop policy π
  - State process $X^{i,N}$ of agent i

$$\begin{cases} X_0^{i,N} & = & \xi(\Gamma^i) \\ X_{t+1}^{i,N} & = & F\big(X_t^{i,N}, \alpha_t^{\pi,i}, \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}, \frac{1}{N}\sum_{j=1}^N \delta_{\alpha_t^{\pi,j}}, \varepsilon_{t+1}^i\big). \end{cases}$$

Mean-field control problem in the N-population model

- **Mean-field controlled dynamics**:
  - Same initial state function $\xi \in L(I;\mathcal{X})$, and same open-loop policy π
  - State process $X^{i,N}$ of agent i

$$\begin{cases} X_0^{i,N} & = & \xi(\Gamma^i) \\ X_{t+1}^{i,N} & = & F\big(X_t^{i,N}, \alpha_t^{\pi,i}, \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}, \frac{1}{N}\sum_{j=1}^N \delta_{\alpha_t^{\pi,j}}, \varepsilon_{t+1}^i\big). \end{cases}$$

- Gain functional identical for each agent i by indistinguishability:

$$V_N^\pi(\xi) \quad = \quad \mathbb{E}\Big[\sum_{t=0}^\infty \beta^t f\big(X_t^{i,N}, \alpha_t^{\pi,i}, \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}, \frac{1}{N}\sum_{j=1}^N \delta_{\alpha_t^{\pi,j}}\big)\Big]$$

Mean-field control problem in the N-population model

- **Mean-field controlled dynamics**:
  - Same initial state function $\xi \in L(I;\mathcal{X})$, and same open-loop policy π
  - State process $X^{i,N}$ of agent i

$$\begin{cases} X_0^{i,N} & = & \xi(\Gamma^i) \\ X_{t+1}^{i,N} & = & F\Big(X_t^{i,N}, \alpha_t^{\pi,i}, \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}, \frac{1}{N}\sum_{j=1}^N \delta_{\alpha_t^{\pi,j}}, \varepsilon_{t+1}^i\Big). \end{cases}$$

- Gain functional identical for each agent i by indistinguishability:

$$V_N^\pi(\xi) \quad = \quad \mathbb{E}\Big[\sum_{t=0}^\infty \beta^t f\Big(X_t^{i,N}, \alpha_t^{\pi,i}, \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}, \frac{1}{N}\sum_{j=1}^N \delta_{\alpha_t^{\pi,j}}\Big)\Big]$$

▶ Optimal gain for the center of decision (social planner/influencer):

$$V_N(\xi) \quad = \quad \sup_{\pi \in \Pi_{OL}^r} V_N^\pi(\xi).$$

Limiting problem ($N \to \infty$): MKV-MDP

- **Mean-field controlled dynamics**:
  - Initial state function $\xi \in L(I; \mathcal{X})$, and open-loop policy π

Limiting problem ($N \to \infty$): MKV-MDP

- **Mean-field controlled dynamics**:
  - Initial state function $\xi \in L(I; \mathcal{X})$, and open-loop policy π
  - State process X^i of agent i in the infinite population

$$\begin{cases} X_0^i &= \xi(\Gamma^i) \\ X_{t+1}^i &= F\big(X_t^i, \alpha_t^{\pi,i}, \mathbb{P}^0_{X_t^i}, \mathbb{P}^0_{\alpha_t^{\pi,i}}, \varepsilon_{t+1}^i\big). \end{cases}$$

Limiting problem ($N \to \infty$): MKV-MDP

- **Mean-field controlled dynamics**:
  - Initial state function $\xi \in L(I;\mathcal{X})$, and open-loop policy π
  - State process X^i of agent i in the infinite population

$$\begin{cases} X_0^i & = & \xi(\Gamma^i) \\ X_{t+1}^i & = & F\big(X_t^i, \alpha_t^{\pi,i}, \mathbb{P}^0_{X_t^i}, \mathbb{P}^0_{\alpha_t^{\pi,i}}, \varepsilon_{t+1}^i\big). \end{cases}$$

- Gain functional identical for each agent i by indistinguishability:

$$V^\pi(\xi) \quad = \quad \mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t f\big(X_t^i, \alpha_t^{\pi,i}, \mathbb{P}^0_{X_t^i}, \mathbb{P}^0_{\alpha_t^{\pi,i}}\big)\Big]$$

Limiting problem ($N \to \infty$): MKV-MDP

- **Mean-field controlled dynamics**:
  - Initial state function $\xi \in L(I; \mathcal{X})$, and open-loop policy π
  - State process X^i of agent i in the infinite population

$$\begin{cases} X_0^i & = & \xi(\Gamma^i) \\ X_{t+1}^i & = & F\big(X_t^i, \alpha_t^{\pi,i}, \mathbb{P}^0_{X_t^i}, \mathbb{P}^0_{\alpha_t^{\pi,i}}, \varepsilon_{t+1}^i\big). \end{cases}$$

- Gain functional identical for each agent i by indistinguishability:

$$V^\pi(\xi) \quad = \quad \mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t f\big(X_t^i, \alpha_t^{\pi,i}, \mathbb{P}^0_{X_t^i}, \mathbb{P}^0_{\alpha_t^{\pi,i}}\big)\Big]$$

▶ Optimal gain for the center of decision (social planner/influencer):

$$V(\xi) \quad = \quad \sup_{\pi \in \Pi_{OL}^r} V^\pi(\xi).$$

Convergence of the N-agents model

- Propagation of chaos: convergence of the mean-field dynamics

Convergence of the N-agents model

- Propagation of chaos: convergence of the mean-field dynamics

Proposition

Under suitable continuity assumptions on F, given ξ and π, we have for all $i \in \mathbb{N}$, $t \in \mathbb{N}$,

$$X_t^{i,N} \xrightarrow[N\to\infty]{a.s.} X_t^i, \quad d_W\left(\frac{1}{N}\sum_{j=1}^{N}\delta_{X_t^{j,N}}, \mathbb{P}^0_{X_t^i}\right) \xrightarrow[N\to\infty]{a.s.} 0.$$

Convergence of the N-agents model

- Propagation of chaos: convergence of the mean-field dynamics

Proposition

Under suitable continuity assumptions on F, given ξ and π, we have for all $i \in \mathbb{N}$, $t \in \mathbb{N}$,

$$X_t^{i,N} \xrightarrow[N\to\infty]{a.s.} X_t^i, \quad d_W\left(\frac{1}{N}\sum_{j=1}^{N} \delta_{X_t^{j,N}}, \mathbb{P}^0_{X_t^i}\right) \xrightarrow[N\to\infty]{a.s.} 0.$$

- Convergence towards MKV-MDP

Convergence of the N-agents model

- Propagation of chaos: convergence of the mean-field dynamics

Proposition

Under suitable continuity assumptions on F, given ξ and π, we have for all $i \in \mathbb{N}$, $t \in \mathbb{N}$,

$$X_t^{i,N} \xrightarrow[N\to\infty]{a.s.} X_t^i, \quad d_W\left(\frac{1}{N}\sum_{j=1}^{N}\delta_{X_t^{j,N}}, \mathbb{P}^0_{X_t^i}\right) \xrightarrow[N\to\infty]{a.s.} 0.$$

- Convergence towards MKV-MDP

Theorem

Under Lipschitz condition on f, we have for all initial state function ξ:

$$\sup_{\pi\in\Pi_{OL}} |V_N^\pi(\xi) - V^\pi(\xi)| \underset{N\to\infty}{\longrightarrow} 0.$$

Consequently, $V_N \underset{N\to\infty}{\longrightarrow} V$ and any ε-optimal strategy for the limit model is an ε-optimal strategy for the N-individual model for N large enough

Studying V

Proposition

F Lipschitz $\Rightarrow V^\pi$ uniformly continuous, uniformly on $\pi \Rightarrow V$ unif. cont.

Studying V

Proposition

F Lipschitz $\Rightarrow$ V^π uniformly continuous, uniformly on π $\Rightarrow$ V unif. cont.

Hyp. A (randomization): Exists i.i.d. uniform $(U_t(\Gamma))_{t\in\mathbb{N}}$ indep of $\xi(\Gamma)$.
(equivalent : only one)

Studying V

Proposition

F Lipschitz $\Rightarrow V^\pi$ uniformly continuous, uniformly on $\pi \Rightarrow V$ unif. cont.

Hyp. A (randomization): Exists i.i.d. uniform $(U_t(\Gamma))_{t\in\mathbb{N}}$ indep of $\xi(\Gamma)$.
(equivalent : only one)

Proposition (Law-invariance)

For any ξ, $\widetilde{\xi} \in L(I;\mathcal{X})$ satisfying **A**, *s.t.* $\mathbb{P}_\xi = \mathbb{P}_{\widetilde{\xi}}$, *we have*

$$V(\xi) \quad = \quad V(\widetilde{\xi}).$$

Theorem (Dynamic Programming Principle)

For any $\mu \in \mathcal{P}(\mathcal{X})$,

$$V(\mu) = \sup_{a \in L(\mathcal{X} \times [0,1]; A)} \mathbb{E}\Big[f\big(\xi, a(\xi, U), \mu, \mathbb{P}_{a(\xi,U)}\big) + \beta V(\mathbb{P}^0_{X_1^{\xi,a}}) \Big]$$

Theorem (Dynamic Programming Principle)

For any $\mu \in \mathcal{P}(\mathcal{X})$,

$$V(\mu) = \sup_{a \in L(\mathcal{X}\times[0,1];A)} \mathbb{E}\Big[f\big(\xi, a(\xi, U), \mu, \mathbb{P}_{a(\xi,U)}\big) + \beta V(\mathbb{P}^0_{X_1^{\xi,a}})\Big]$$

Useful to study V:

Theorem (Dynamic Programming Principle)

For any $\mu \in \mathcal{P}(\mathcal{X})$,

$$V(\mu) = \sup_{a \in L(\mathcal{X}\times[0,1];A)} \mathbb{E}\Big[f\big(\xi, a(\xi,U), \mu, \mathbb{P}_{a(\xi,U)}\big) + \beta V(\mathbb{P}^0_{X_1^{\xi,a}})\Big]$$

Useful to study V:

- Convexity?
- Lipschitz?

Theorem (Dynamic Programming Principle)

For any $\mu \in \mathcal{P}(\mathcal{X})$,

$$V(\mu) = \sup_{a \in L(\mathcal{X} \times [0,1]; A)} \mathbb{E}\Big[f\big(\xi, a(\xi, U), \mu, \mathbb{P}_{a(\xi,U)}\big) + \beta V(\mathbb{P}^0_{X_1^{\xi,a}})\Big]$$

Useful to study V:

- Convexity?
- Lipschitz?

max reached measurably in $\mu \Rightarrow$ stationary feedback policy.
Issues:

Theorem (Dynamic Programming Principle)

For any $\mu \in \mathcal{P}(\mathcal{X})$,

$$V(\mu) = \sup_{a \in L(\mathcal{X}\times[0,1];A)} \mathbb{E}\Big[f\big(\xi, a(\xi, U), \mu, \mathbb{P}_{a(\xi,U)}\big) + \beta V(\mathbb{P}^0_{X_1^{\xi,a}})\Big]$$

Useful to study V:

- Convexity?
- Lipschitz?

max reached measurably in $\mu \Rightarrow$ stationary feedback policy.
Issues:

- sup on an ∞-dimensional space + iterations.

Theorem (Dynamic Programming Principle)

For any $\mu \in \mathcal{P}(\mathcal{X})$,

$$V(\mu) \quad = \quad \sup_{a \in L(\mathcal{X} \times [0,1]; A)} \mathbb{E}\Big[f\big(\xi, a(\xi, U), \mu, \mathbb{P}_{a(\xi,U)}\big) + \beta V(\mathbb{P}^0_{X_1^{\xi,a}})\Big]$$

Useful to study V:

- Convexity?
- Lipschitz?

max reached measurably in $\mu \Rightarrow$ stationary feedback policy.
Issues:

- sup on an ∞-dimensional space + iterations.
- Hard to store policies.

Theorem (Dynamic Programming Principle)

For any $\mu \in \mathcal{P}(\mathcal{X})$,

$$V(\mu) = \sup_{a \in L(\mathcal{X} \times [0,1]; A)} \mathbb{E}\Big[f\big(\xi, a(\xi, U), \mu, \mathbb{P}_{a(\xi,U)}\big) + \beta V(\mathbb{P}^0_{X_1^{\xi,a}})\Big]$$

Useful to study V:

- Convexity?
- Lipschitz?

max reached measurably in $\mu \Rightarrow$ stationary feedback policy.
Issues:

- sup on an ∞-dimensional space + iterations.
- Hard to store policies.
- Hard to compute X.

Proposition

$\mathcal{X}$, $\mathcal{A}$ finite $\Rightarrow$ equivalent to MDP on state/action spaces $[0,1]^n$.

Proposition

$\mathcal{X}$, $\mathcal{A}$ *finite* $\Rightarrow$ *equivalent to MDP on state/action spaces* $[0,1]^n$.

$$\begin{aligned} X_t &\to (\mathbb{P}(X_t = x))_{x \in \mathcal{X}} \\ \alpha_t &\to (\mathbb{P}(\alpha_t = a \mid X_t = x))_{a \in A, x \in \mathcal{X}} \end{aligned}$$

Proposition

$\mathcal{X}$, $\mathcal{A}$ finite $\Rightarrow$ equivalent to MDP on state/action spaces $[0,1]^n$.

$$\begin{array}{ccl} X_t & \to & (\mathbb{P}(X_t = x))_{x \in \mathcal{X}} \\ \alpha_t & \to & (\mathbb{P}(\alpha_t = a \mid X_t = x))_{a \in A, x \in \mathcal{X}} \end{array}$$

- Dynamic inherits Lipschitz property from F $\to$ MDP algo applies.
- $(\mathbb{P}_{X_t^a})_t$ = simple vectors in $[0,1]^{|\mathcal{X}|-1}$.
- Easy to compute perfectly
- $\mathcal{X}$, $\mathcal{A}$ not finite $\to$ discretization $\to$ finite spaces.

Back to targeted advertising example: spaces and noise

- State space $\mathcal{X} = \{0, 1\}$:
  - $x = 1$ (resp. 0): customer (not customer) of the company C

Back to targeted advertising example: spaces and noise

- State space $\mathcal{X} = \{0, 1\}$:
  - $x = 1$ (resp. 0): customer (not customer) of the company C
- Action space $A = \{0, 1\}$:
  - $a = 1$ (resp. 0): SN send (or not) an ad

Back to targeted advertising example: spaces and noise

- State space $\mathcal{X} = \{0, 1\}$:
  - $x = 1$ (resp. 0): customer (not customer) of the company C
- Action space $A = \{0, 1\}$:
  - $a = 1$ (resp. 0): SN send (or not) an ad
- For each player i:
  - ε_t^i: uniform r.v. representing e.g. time spent at day t on a forum about phones
- For simplicity, no common noise

Targeted advertising example: dynamics and reward

- State transition function:

$$F(x, \mu, a, e) = \begin{cases} \mathbf{1}_{e > \mu(\{0\}) - \eta a} & \text{if} \quad x = 0 \\ \mathbf{1}_{e < \mu(\{1\}) + \eta a} & \text{if} \quad x = 1. \end{cases}$$

Targeted advertising example: dynamics and reward

- State transition function:

$$F(x,\mu,a,e) = \begin{cases} \mathbf{1}_{e>\mu(\{0\})-\eta a} & \text{if} \quad x=0 \\ \mathbf{1}_{e<\mu(\{1\})+\eta a} & \text{if} \quad x=1. \end{cases}$$

  - Large e: eager to change of phone

Targeted advertising example: dynamics and reward

- State transition function:

$$F(x, \mu, a, e) = \begin{cases} \mathbf{1}_{e > \mu(\{0\}) - \eta a} & \text{if} \quad x = 0 \\ \mathbf{1}_{e < \mu(\{1\}) + \eta a} & \text{if} \quad x = 1. \end{cases}$$

  - Large e: eager to change of phone
  - $\mu(\{0\})$: proportion of SN users that are not customers of C

Targeted advertising example: dynamics and reward

- State transition function:

$$F(x,\mu,a,e) \;=\; \begin{cases} \mathbf{1}_{e>\mu(\{0\})-\eta a} & \text{if} \quad x=0 \\ \mathbf{1}_{e<\mu(\{1\})+\eta a} & \text{if} \quad x=1. \end{cases}$$

  - Large e: eager to change of phone
  - $\mu(\{0\})$: proportion of SN users that are not customers of C
  - $\eta > 0$: impact of ad for incentive to become or remain a customer of C

Targeted advertising example: dynamics and reward

- State transition function:

$$F(x,\mu,a,e) = \begin{cases} \mathbf{1}_{e>\mu(\{0\})-\eta a} & \text{if} \quad x=0 \\ \mathbf{1}_{e<\mu(\{1\})+\eta a} & \text{if} \quad x=1. \end{cases}$$

  - Large e: eager to change of phone
  - $\mu(\{0\})$: proportion of SN users that are not customers of C
  - $\eta > 0$: impact of ad for incentive to become or remain a customer of C

- Reward function: for $x \in \mathcal{X} = \{0,1\}, a \in A = \{0,1\}$,

$$f(x,a) = x - ca,$$

  - $c > 0$: ad cost

Reformulation as MDP on $[0,1]$

- $\mathcal{P}(\mathcal{X}) \leftrightarrow$ Bernoulli parameter $p \in [0,1]$
- $\mathcal{P}(A) \leftrightarrow$ Bernoulli parameter $q \in [0,1]$
- Policy $\pi \rightarrow p_t^\pi$: Bernoulli parameter of $\mathbb{P}_{X_t}$, i.e. $p_t^\pi = \mathbb{E}[X_t]$:

$$p_{t+1}^\pi \;=\; \Phi(p_t^\pi, q_t^\pi) \;:=\; p_t^\pi + q_t^\pi \, \min(\eta, 1 - p_t^\pi), \quad t \in \mathbb{N},$$

where

- $q_t^\pi = \mathbb{E}[\alpha_t^\pi]$: probability of displaying an ad $\equiv$ sending an ad to q_t^π proportion of the SN users

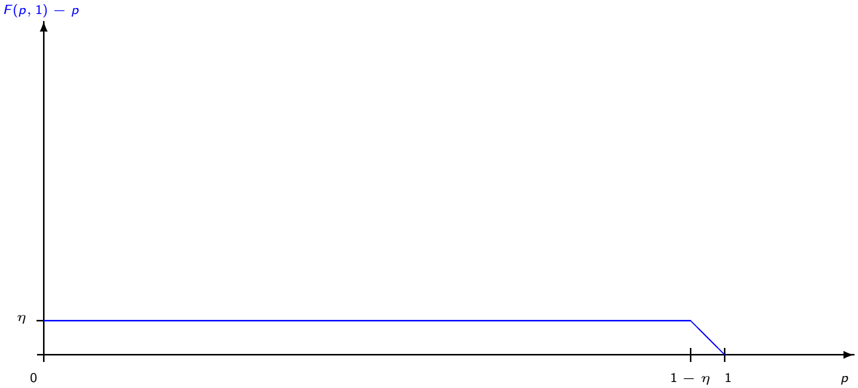
▶ Value function on $[0,1]$:

$$V(p) \;=\; \sup_{q_t^\pi \in [0,1]} \sum_{t=0}^{\infty} \beta^t (p_t^\pi - c q_t^\pi), \quad p \in [0,1],$$

satisfying the DP:

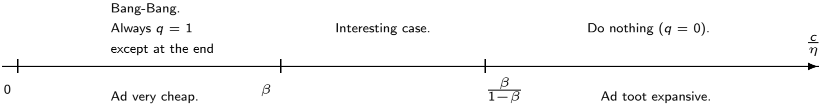
$$V(p) \;=\; \sup_{q \in [0,1]} \Big[p - cq + \beta V\big(p + \min(\eta, 1-p)\big) \Big]$$

Variational graph



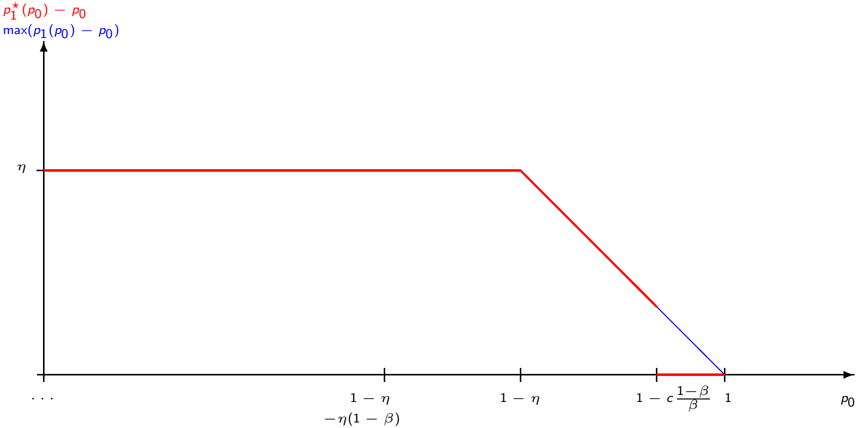
Maximal variations $F(p,1) - p$ (in blue).

Disjunction



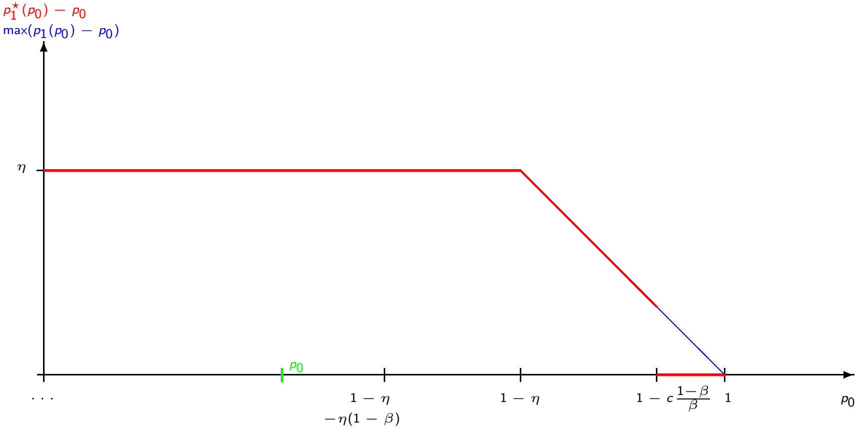
Disjunction according to the position of $\frac{c}{\eta}$ relative to β and $\frac{\beta}{1-\beta}$.

Optimal variation



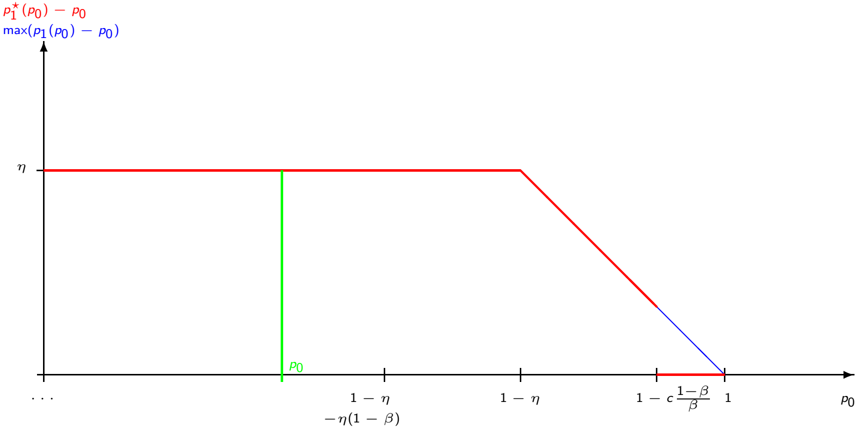
Case $\frac{c}{\eta} < \beta$. Variational optimal policy (in red).

Optimal variation



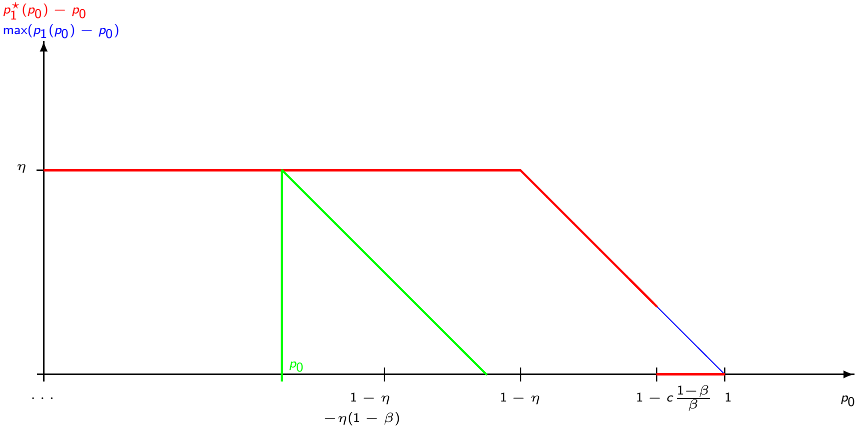
Case $\frac{c}{\eta} < \beta$. Variational optimal policy (in red).

Optimal variation



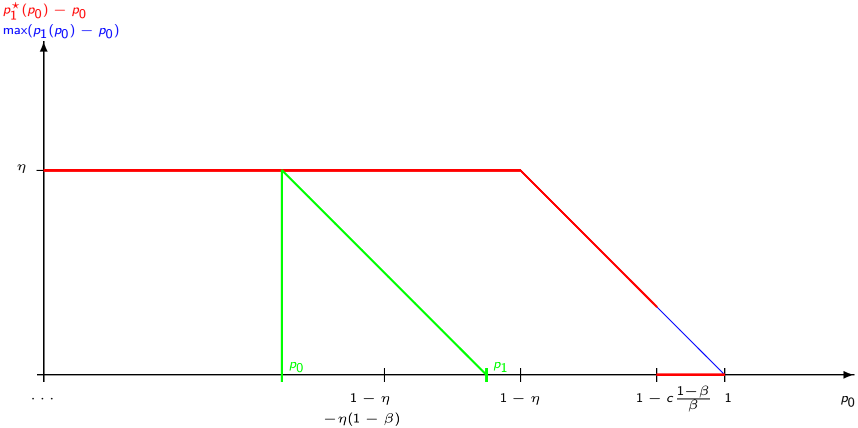
Case $\frac{c}{\eta} < \beta$. Variational optimal policy (in red).

Optimal variation



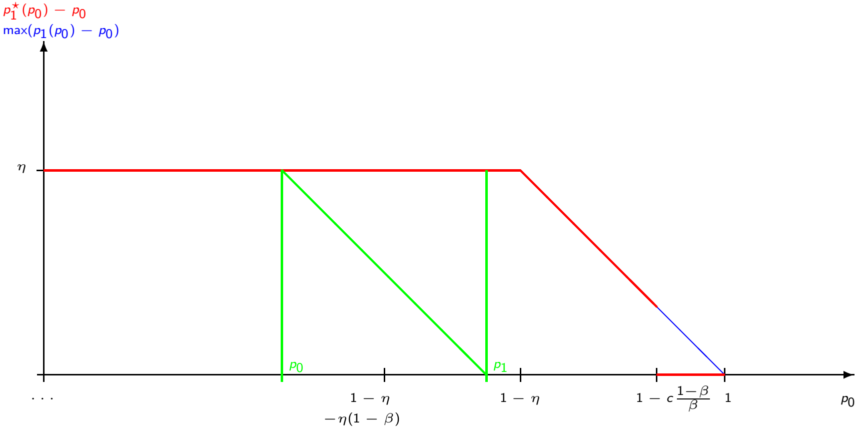
Case $\frac{c}{\eta} < \beta$. Variational optimal policy (in red).

Optimal variation



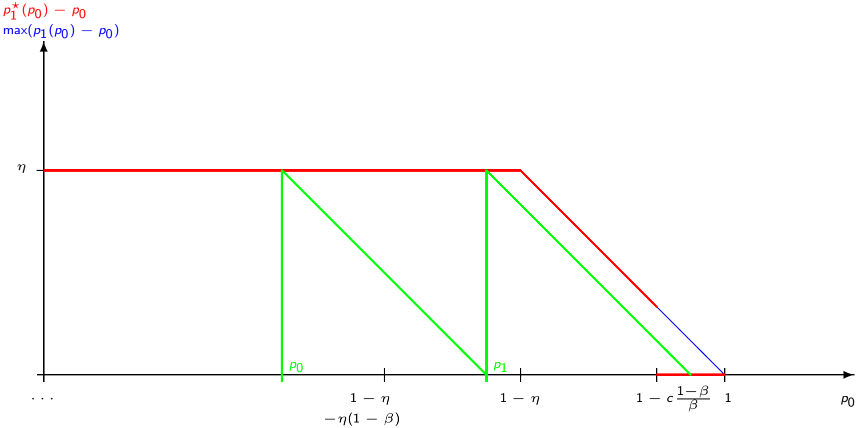
Case $\frac{c}{\eta} < \beta$. Variational optimal policy (in red).

Optimal variation



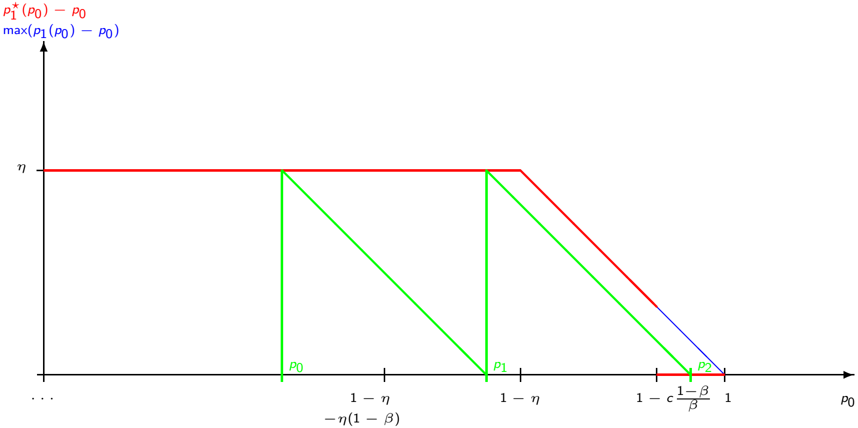
Case $\frac{c}{\eta} < \beta$. Variational optimal policy (in red).

Optimal variation



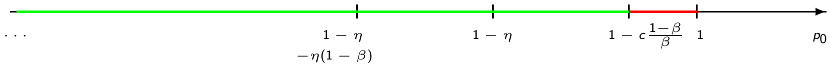
Case $\frac{c}{\eta} < \beta$. Variational optimal policy (in red).

Optimal variation



Case $\frac{c}{\eta} < \beta$. Variational optimal policy (in red).

Optimal variation

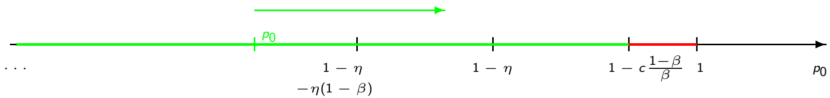


Case $\frac{c}{\eta} < \beta$. Variational optimal policy, Bang-Bang.

Optimal variation

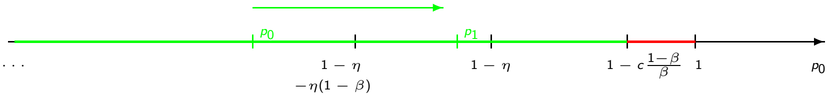
Case $\frac{c}{\eta} < \beta$. Variational optimal policy, Bang-Bang.

Optimal variation



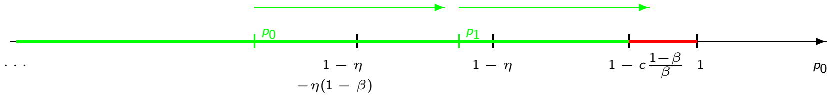
Case $\frac{c}{\eta} < \beta$. Variational optimal policy, Bang-Bang.

Optimal variation



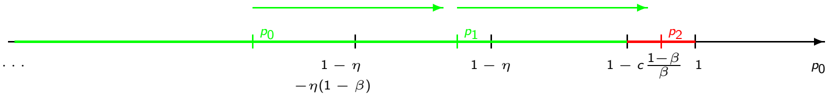
Case $\frac{c}{\eta} < \beta$. Variational optimal policy, Bang-Bang.

Optimal variation



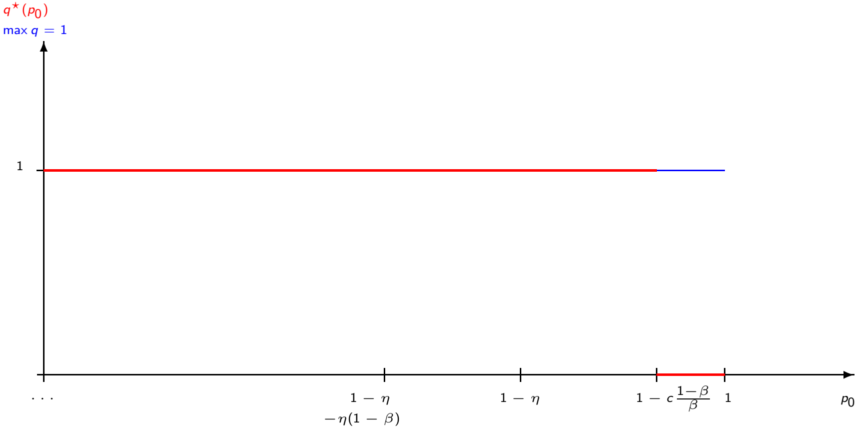
Case $\frac{c}{\eta} < \beta$. Variational optimal policy, Bang-Bang.

Optimal variation



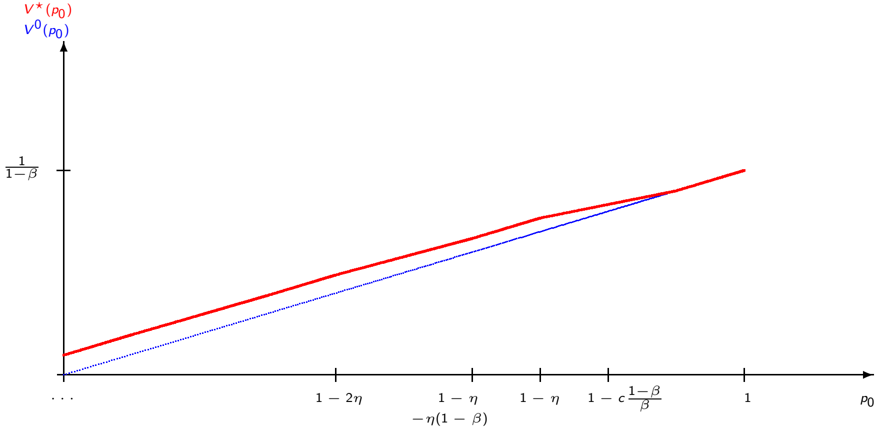
Case $\frac{c}{\eta} < \beta$. Variational optimal policy, Bang-Bang.

Optimal policy



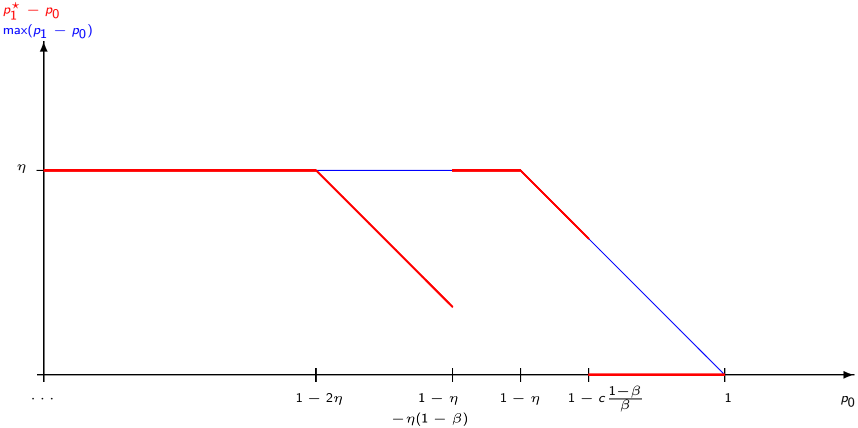
Case $\frac{c}{\eta} < \beta$. Optimal policy (in red).

Optimal value



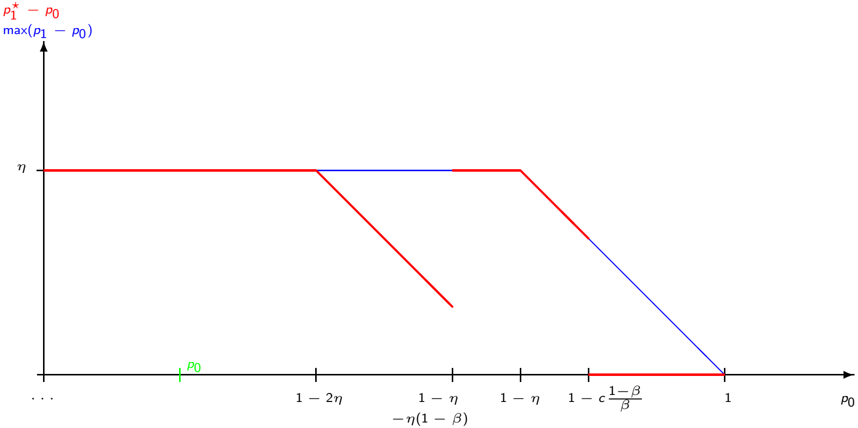
Case $\frac{c}{\eta} < \beta$. Optimal value (in red) against "no ad" value (in blue).

Optimal variation



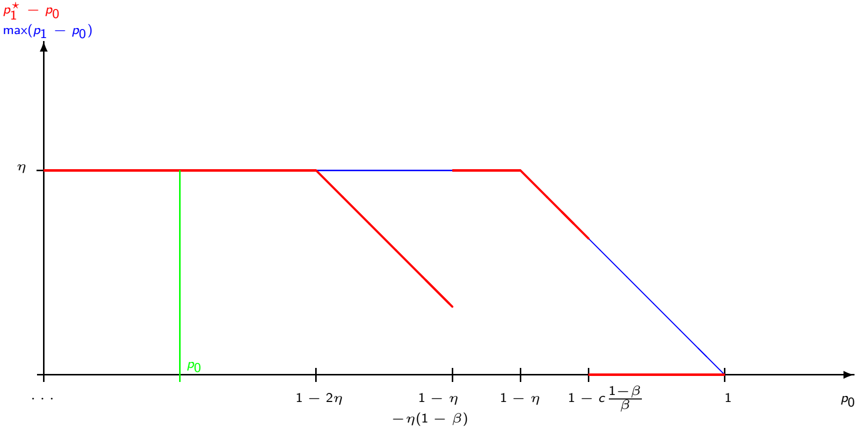
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



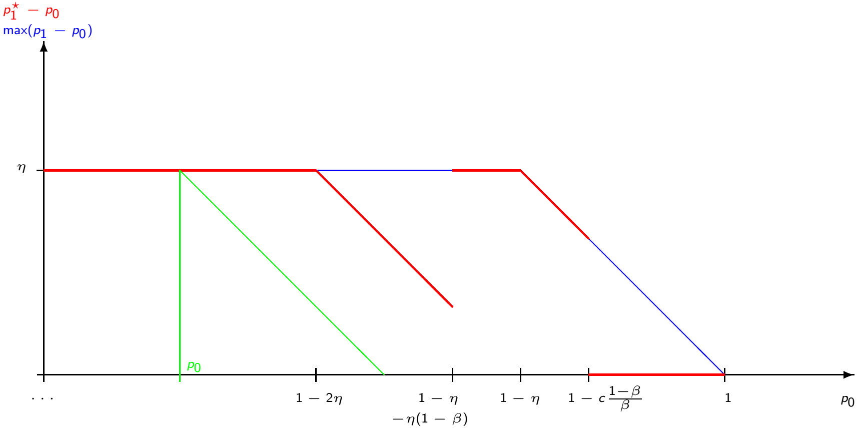
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



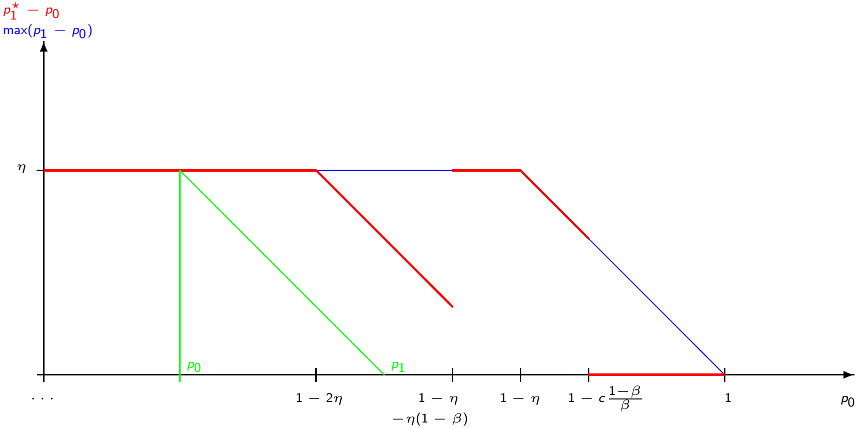
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



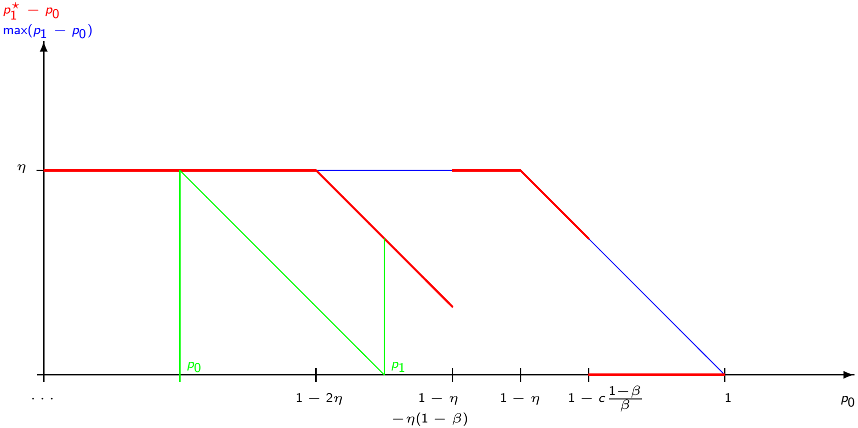
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



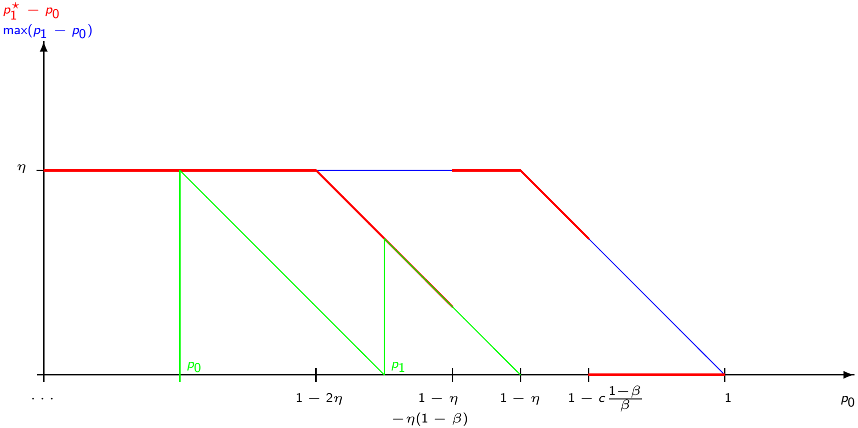
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



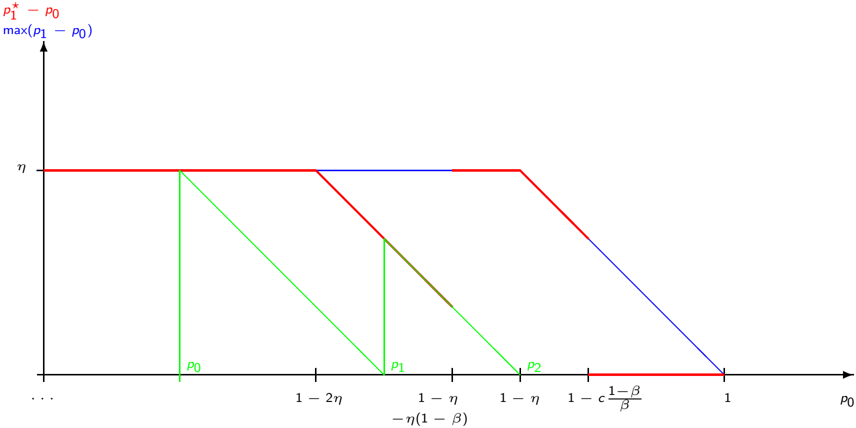
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



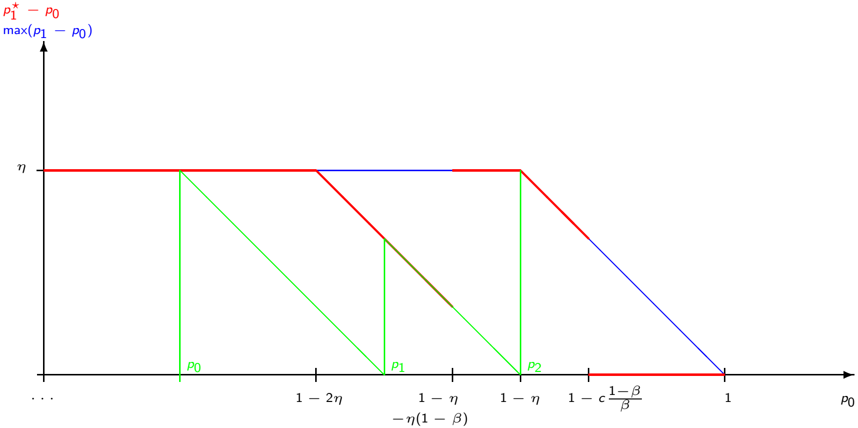
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



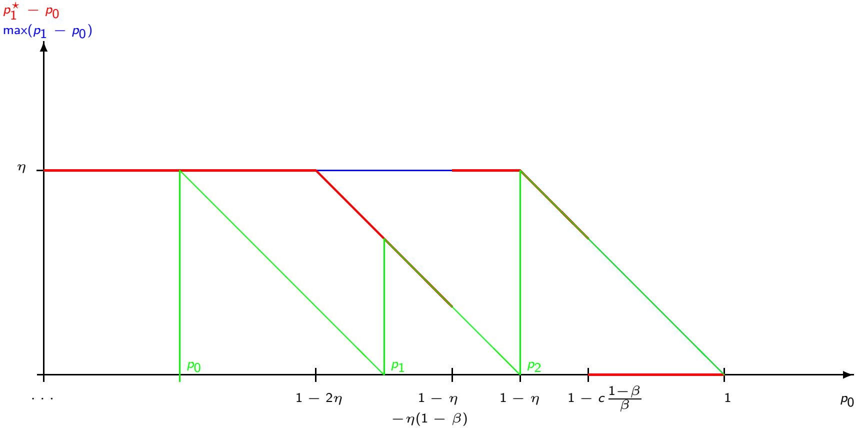
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



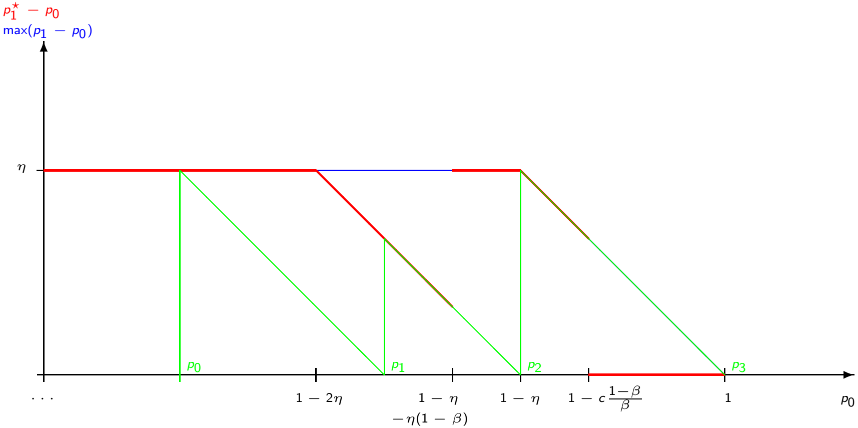
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



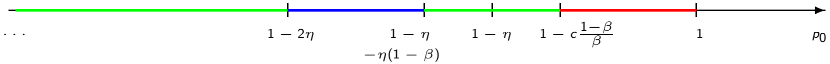
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



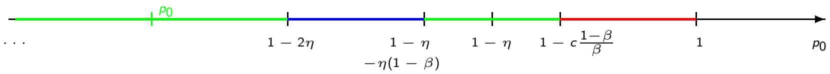
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy (in red).

Optimal variation



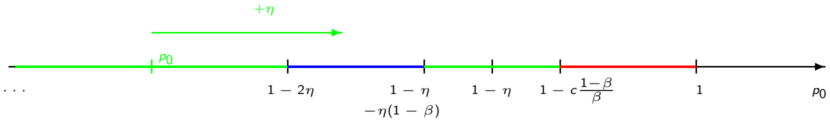
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1-\eta$" case.

Optimal variation



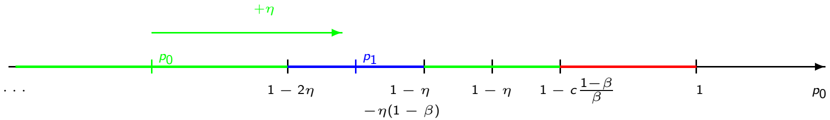
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1-\eta$" case.

Optimal variation



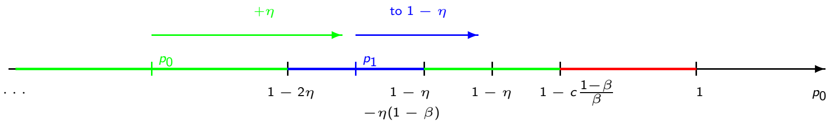
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1-\eta$" case.

Optimal variation



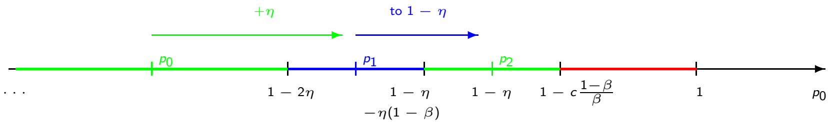
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1-\eta$" case.

Optimal variation



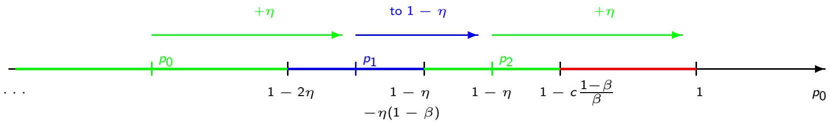
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1-\eta$" case.

Optimal variation



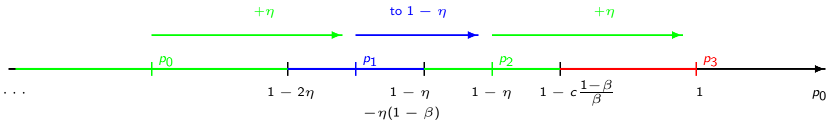
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1-\eta$" case.

Optimal variation



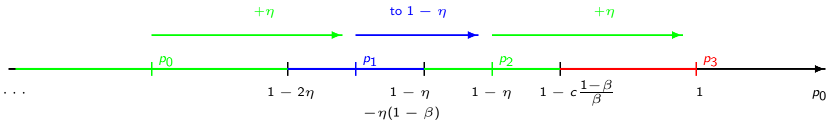
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1-\eta$" case.

Optimal variation



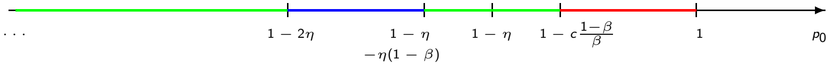
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1 - \eta$" case.

Optimal variation



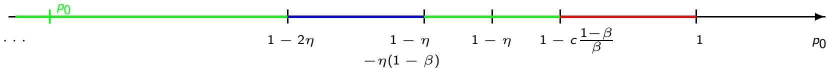
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, "jump to $1-\eta$" case.

Optimal variation



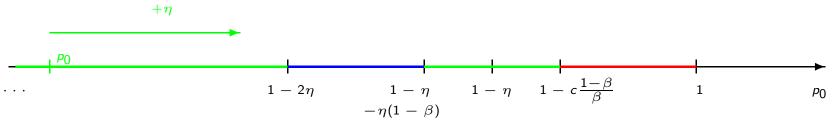
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal variation



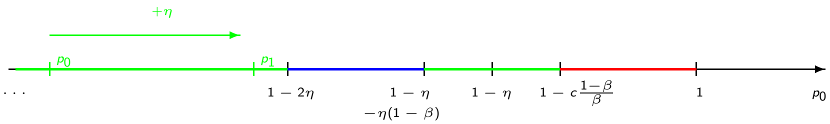
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal variation



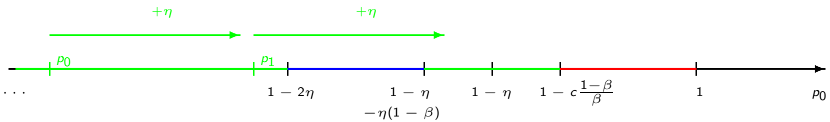
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal variation



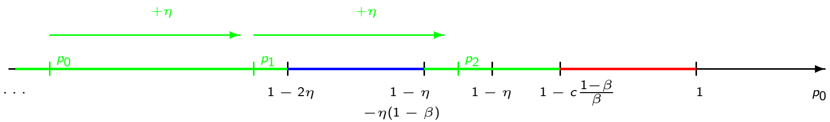
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal variation



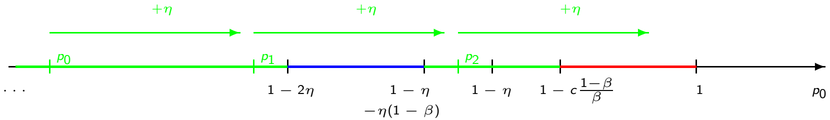
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal variation



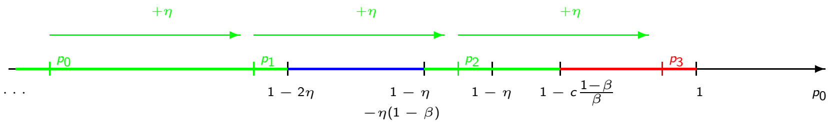
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal variation



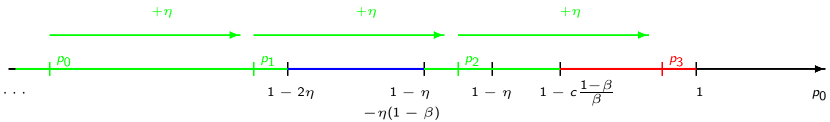
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal variation



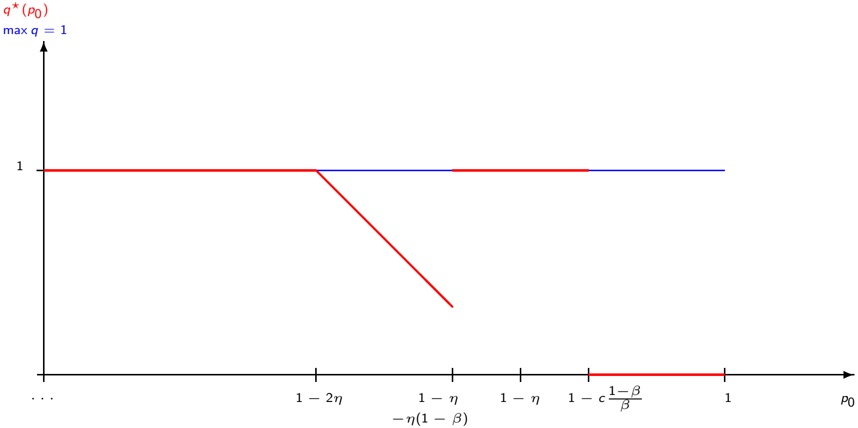
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal variation



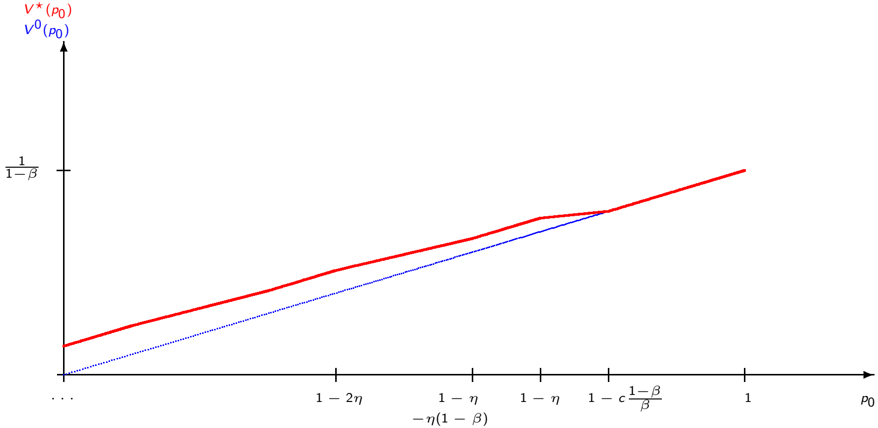
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Variational optimal policy, Bang-Bang case.

Optimal control



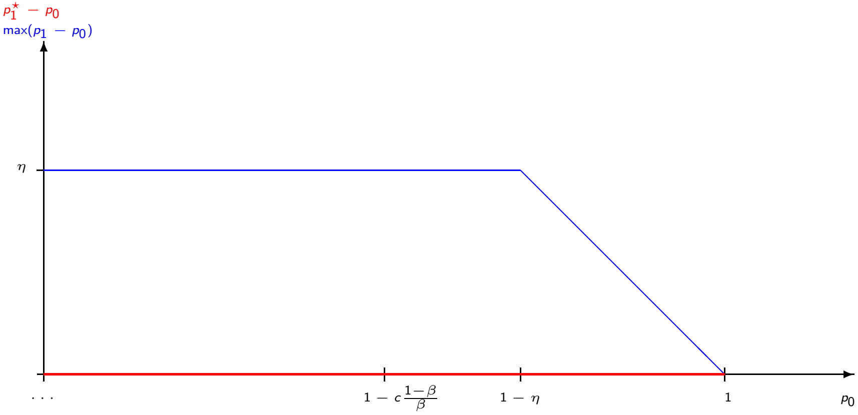
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Optimal policy (in red).

Optimal policy



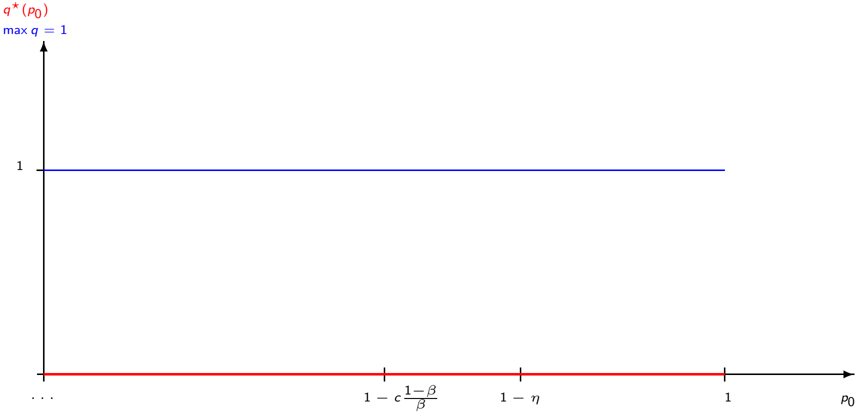
Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Optimal value (in red) against "no ad" value (in blue).

Optimal variation



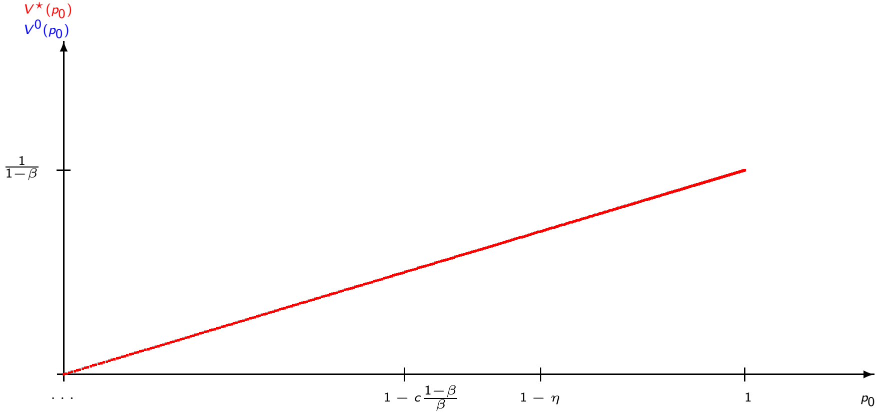
Case $\frac{\beta}{1-\beta} < \frac{c}{\eta}$. Variational optimal policy (in red).

Optimal control



Case $\frac{\beta}{1-\beta} < \frac{c}{\eta}$. Optimal policy (in red).

Optimal value



Case $\beta < \frac{c}{\eta} < \frac{\beta}{1-\beta}$. Optimal value (in red) against "no ad" value (in blue).

Sketch of the proof

Let $(p_t)_{t\in\mathbb{N}}$ = state process from optimal control.
Group rewards by increments

$$\Rightarrow V(p_0) = \frac{p_0}{1-\beta} + \sum_{t\in\mathbb{N}} \beta^t (p_{t+1} - p_t) \left(\frac{\beta}{1-\beta} - \frac{c}{\min(\eta, 1-p_t)} \right)$$

If $\frac{\beta}{1-\beta} - \frac{c}{\eta} \leqslant 0$, $V(p_0) \leqslant \frac{p_0}{1-\beta} = V^0(p_0) \Rightarrow$ no ad.
If $\frac{\beta}{1-\beta} - \frac{c}{\eta} > 0$:

- If $p_t \geqslant 1 - c\frac{1-\beta}{\beta}$, then $\frac{\beta}{1-\beta} - \frac{c}{\min(\eta,1-p_t)} \leqslant 0 \Rightarrow$ no ad after p_t.
- If $p_t \in [1-\eta, 1 - c\frac{1-\beta}{\beta}[$, then

$$\begin{aligned} &\sum\nolimits_{s \geqslant t} \beta^s (p_{s+1} - p_s) \left(\frac{\beta}{1-\beta} - \frac{c}{\min(\eta,1-p_s)} \right) \\ &\leqslant \beta^t (1 - p_t) \left(\frac{\beta}{1-\beta} - \frac{c}{\min(\eta,1-p_t)} \right) \end{aligned}$$

$\Rightarrow q_t = 1$ and $q_s = 0$ for $s > t$.

Sketch of the proof

Still in the case $\frac{\beta}{1-\beta} - \frac{c}{\eta} > 0$:
If $p_t \leqslant 1 - \eta$:
assuming $p_{t+1} \leqslant 1 - \eta$,
gain between p_t and p_{t+2}:

$$p_t - c\frac{p_{t+1} - p_t}{\eta} + \beta\left(p_{t+1} - c\frac{p_{t+2} - p_{t+1}}{\eta}\right)$$

Derivative in p_{t+1}: $-\frac{c}{\eta} + \beta(1 + \frac{c}{\eta}) = (1-\beta)(\frac{\beta}{1-\beta} - \frac{c}{\eta}) > 0$.
$\Rightarrow$ p_{t+1} can't be moved to the right (would contradict optimality).
Implies

- if $p_t \leqslant 1 - 2\eta$, $p_{t+1} = p_t + \eta$.
- if $1 - 2\eta < p_t \leqslant 1 - \eta$, $p_{t+1} \geqslant 1 - \eta$

Sketch of the proof

Only thing left:
If $1-2\eta < p_t \leqslant 1-\eta$, where is p_{t+1} in $[1-\eta, p_t+\eta]$?
Notice: p_{t+1} known $\Rightarrow$ rest of the trajectory known $\Rightarrow$ $V(p_{t+1})$ known.
DPP $\Rightarrow$ Maximize $p_t - c\frac{p_{t+1}-p_t}{\eta} + \beta V(p_{t+1})$ over $p_{t+1} \in [1-\eta, p_t+\eta] \Rightarrow$ Simple.
$\Rightarrow$ Disjunction between $\frac{c}{\eta} > \beta$ and $\frac{c}{\eta} \leqslant \beta$.

Variant: fast targeted advertising

Issue with previous application: Control happens too late
Nowadays: ad displayed between forum reading and choice

Variant: fast targeted advertising

Issue with previous application: Control happens too late
Nowadays: ad displayed between forum reading and choice
$\Rightarrow \alpha_t$ should depend on ε_{t+1}

Variant: fast targeted advertising

Issue with previous application: Control happens too late
Nowadays: ad displayed between forum reading and choice
$\Rightarrow \alpha_t$ should depend on ε_{t+1}
Problem: not adapted!

Variant: fast targeted advertising

Issue with previous application: Control happens too late
Nowadays: ad displayed between forum reading and choice
$\Rightarrow \alpha_t$ should depend on ε_{t+1}
Problem: not adapted!
Solution: α_t valued in $A^{[0,1]}$

Variant: fast targeted advertising

Issue with previous application: Control happens too late
Nowadays: ad displayed between forum reading and choice
$\Rightarrow$ α_t should depend on ε_{t+1}
Problem: not adapted!
Solution: α_t valued in $A^{[0,1]}$
$\Rightarrow$ Explicit solution: "Bang-Bang" control.

- $c < \frac{\beta}{1-\beta}$ $\rightarrow$ send ad to all "hesitating" individual.
- $c \geqslant \frac{\beta}{1-\beta}$ $\rightarrow$ send no ad.

Thank you for your attention!