# An Actor-Critic Contextual Bandit Algorithm for Personalized Interventions using Mobile Devices

Huitian Lei

Joint work with Ambuj Tewari and Susan Murphy
Department of Statistics
University of Michigan

August 9, 2015

# Background

- Mobile devices such as smartphones and wearable devices serve a great platform to deliver health interventions.
- Adaptive intervention (AI, or Dynamic treatment regime, treatment policy)
  - A sequence of decision rules/ policies that input patients' characteristics and output recommended interventions.
  - Interventions are (1) personalized, and (2) adapted to patients' ongoing performances.
- Just in time adaptive intervention (JITAI) are *interventions that are delivered in real-time, and are adapted to address the immediate and changing needs of individuals as they go about their daily lives* (Nahum-Shani et al 2014)

# Motivation and Goals

Motivation:

- Lack of guidance to build high quality JITAI despite its increasing popularity.

Glossary: JITAI= treatment policy

Goals:

- An online learning algorithm that learns the optimal treatment policy.
- Make statistical inference on the optimal policy. Address important scientific question:
  - which patient variable is useful for adapting the intervention?

# Problem Formulation: Background on Bandit problems

The stochastic multi-armed bandit problem:

- A gambler (decision maker) needs to choose from multiple arms.
- Each arm is associated with an unknown distribution of rewards.
- His goal is to maximize the expected sum of rewards.
- Carefully tradeoff between acquiring new knowledge (called "exploration") and optimizing his decisions based on existing knowledge (called "exploitation")

Contextual bandits: bandits with covariates, bandits with side information.

# Problem Formulation: Contextual Bandit, Notations

- A sequence of decision points: $t = 1, 2, ....$
- Context space: $\mathcal{S}$.
- Discrete action space: $\mathcal{A}$.
- At decision point $t$, the decision maker observes $S_t \in \mathcal{S}$ and make action $A_t \in \mathcal{A}$, a reward $R_t$ is revealed before the next decision point.
- Contextual bandit assumes that the sequence of contexts $\{S_t\}_{t=1}^T$ are i.i.d. with some fixed unknown distribution $d(s)$: $A_t$ affects $R_t$, not the distribution of $S_{t+1}$.
- Linear expected reward: $\mathbb{E}(R|S = s, A = a) = f(s, a)^T \mu^*$. $f(s, a)$ is the reward feature.

# The Optimal Policy

A class of parametrized stationary policies : $\pi_\theta(a|s)$. In particular we focus on logistic policies: $\pi_\theta(a|s) = \frac{\exp\{\theta^T g(s,a)\}}{\sum_{b\in\mathcal{A}} \exp\{\theta^T g(s,b)\}}$, $\theta \in \Theta$. $g(s,a)$ is a low dimensional policy feature.
The average reward of policy $\pi_\theta$ is

$$V(\theta, \mu^*) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} f(s,a)^T \mu^* \pi_\theta(a|s)$$

1. When $E(R|S=s, A=a)$ is a constant for all $a$, maximizing $V(\theta)$ is an ill-posed problem since solution is not unique.
2. When $E(R|S=s, A=a)$ is not a constant, it is easy to provide example where the maximizer of $V(\theta)$ has $\theta_i$s equal to $\infty$: the policy that maximizes $V(\theta)$ is deterministic . Deterministic policy leads to habituation and boredom (Epstein et al 2008). Intervention variety is therapeutic.

# The stochasticity constraint and the regularized average reward

Mathematize intervention variety for binary action space $\mathcal{A} = \{0, 1\}$,

- A stochasticity constraint: $1 - \beta \leq P(p_0 \leq \pi_\theta(a = 1|s) \leq 1 - p_0)$.
- A relaxed and smoother stochasticity constraint by applying Markov inequality. $g(s) = g(s, 1) - g(s, 0)$

$$\theta^T \mathbb{E}[g(s)g(s)^T]\theta \leq (\log(\frac{p_0}{1 - p_0}))^2 \beta$$

The regularized average reward

$$J(\theta, \mu^*) = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} f(s, a)^T \mu^* \pi_\theta(s, a) - \lambda \theta^T \mathbb{E}[g(s)g(s)^T]\theta$$

The optimal policy parameter

$$\theta^* = \operatorname*{argmax}_\theta J(\theta, \mu^*)$$

# An Actor-Critic Algorithm: initialization

Initialization:

- $T$ is the total number of decision points.
- $d$ is dimension of $f(s, a)$. $p$ is the dimension of $g(s, a)$.
- Critic initialization: $B(0) = \zeta I_{d\times d}$, $I_{d\times d}$ is a $d \times d$ identity matrix. $A(0) = 0_d$, $\mu_0 = 0_d$, are $d \times 1$ column vectors.
- Actor initialization: $\theta_0$ to be the theory based policy parameter provided by behavioral scientists.

# An Actor-Critic Algorithm: the Actor and the Critic

**Algorithm 1:** An online actor critic algorithm

Start from $t = 0$.

**while** $t < T$ **do**

At decision point $t$, observe context $s_t$ ;

Draw an action $a_t$ according to probability distribution $\pi_{\theta_{t-1}}(s_t, a)$ ;

Observe a reward $r_t$ ;

Critic update:

$B(t) = B(t-1) + f(s_t, a_t) f(s_t, a_t)^T$, $A(t) = A(t-1) + f(s_t, a_t) r_t$,

$\mu_t = B(t)^{-1} A(t)$. ;

Actor update:

$$\theta_t = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{a} f(s_\tau, a)^T \mu_t \pi_\theta(s_\tau, a) - \lambda \theta^T [\frac{1}{t} \sum_{\tau=1}^{t} g(s_\tau, 1)^T g(s_\tau, 1)] \theta$$

Go to decision point $t + 1$.

**end**

# Asymptotic Theory: the Critic

## Theorem

*(Asymptotic properties of the critic) The critic's estimate $\mu_t$ converges to $\mu^*$ in probability. The convergence rate is $O(1/\sqrt{t})$, the optimal parametric convergence rate. Furthermore, $\sqrt{t}(\mu_t - \mu^*)$ converges in distribution to multivariate normal with mean $0_d$ and covariance matrix $[\mathbb{E}_{\theta^*}(f(s,a)f(s,a)^T)]^{-1}\sigma^2$, where*

$$\mathbb{E}_\theta(f(s,a)f(s,a)^T) = \sum_s d(s) \sum_a f(s,a)f(s,a)^T \pi_\theta(s,a)$$

*. The plug-in estimator of the asymptotic covariance is consistent.*

# Asymptotic Theory: the Actor

## Theorem

*(Asymptotic properties of the actor) The actor's estimate $\theta_t$ converges to $\theta^*$ in probability. The convergence rate is $O(1/\sqrt{t})$. Furthermore, $\sqrt{t}(\theta_t - \theta^*)$ converges in distribution to multivariate normal with mean $0_p$ and covariance matrix $(J_{\theta\theta}(\mu^*, \theta^*)^{-1} V^* (J_{\theta\theta}(\mu^*, \theta^*))^{-1}$, where*

$$V^* = \sigma^2 J_{\theta\mu}(\mu^*, \theta^*)\mathbb{E}_\theta(f(s,a)f(s,a)^T J_{\mu\theta}(\mu^*, \theta^*) + \sum_s d(s) j_\theta(\mu^*, \theta^*, s) j_\theta(\mu^*, \theta^*, s)^T$$

*. The plug-in estimator of the asymptotic covariance is consistent.*

In the expression of asymptotic covariance matrix,

$$j(\mu, \theta, s) = \sum_a f(s,a)^T \mu \pi_\theta(s,a) - \lambda \theta^T [g(s)g(s)^T]\theta$$

$$J(\mu, \theta) = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} E(R|s,a)\pi_\theta(s,a) - \lambda \theta^T [g(s)g(s)^T]\theta$$

$J_{\theta\theta}$ and $J_{\theta\mu}$ are the second order partial derivatives of $J$. $j_\theta$ is the first order partial derivative of $j$.

# Why relate to high dimensional statistics?

- Reward prediction $E(R|S = s, A = a) = f(s, a)^T \mu^*$ problem usually involves high dimensional reward feature $f(s, a)$.
- Likely to allow $d$ (reward feature dimension) to grow with sample size.

# Thank you

Wald Lecture on:
Continual, Online Learning in Sequential Decision Making
Thursday, August 13, 10:30 a.m.
Susan A. Murphy, University of Michigan

# Creating Bootstrap Confidence Intervals

- Plug in variance estimator very sensitive to the estimated value $\mu_T$ and $\theta_T$. Underestimated variance leads to anti-conservative confidence intervals.
- Creating a bootstrap sample $\{\theta_T^b\}_{b=1}^B$ by boostrapping the residuals $\{\epsilon_t = r_t - f(s_t, a_t)^T \mu_T\}_{t=1}^T$
- Percentile-t bootstrap confidence interval.

# Creating Bootstrap Confidence Intervals

**Algorithm 2:** Generating a bootstrap sample $\theta_T^b$

Start from $t = 0$.
**while** $t < T$ **do**
  Context is $s_t$ ;
  Boostrap an action $a_t^b$ according to probability distribution $\pi_{\theta_{t-1}}(s_t, a)$ ;
  Bootstrap the residuals to generate a bootstrapped reward
  $r_t^b = f(s_t, a_t^b)^T \mu_T + \epsilon_t^b$
  Critic update:
  $\mu_t^b = (\sum_{\tau=1}^T f(s_\tau, a_\tau^b) f(s_\tau, a_\tau^b)^T)^{-1} (\sum_{\tau=1}^T f(s_\tau, a_\tau^b) r_\tau^b)$ ;
  Actor update:

$$\theta_t = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{a} f(s_\tau, a)^T \mu_t^b \pi_\theta(s_\tau, a) - \lambda \theta^T [\frac{1}{t} \sum_{\tau=1}^{t} g(s_\tau, 1)^T g(s_\tau, 1)] \theta$$

  Go to decision point $t + 1$.
**end**