Reinforcement Learning - Lecture 2

Policy Evaluation and Monte-Carlo Methods

ENS M2

October 1, 2025

Contents

1 The Bakery Problem **2**

2 Monte-Carlo Policy Evaluation **2**

2.1 Theoretical Background 2

2.2 Variance Analysis 2

2.3 Variance Bounds 3

2.4 Central Limit Theorem 3

2.5 Variance Estimation 3

3 Non-Asymptotic Bounds **3**

3.1 Bienaymé-Tchebychev Inequality 3

3.2 Hoeffding Inequality 3

4 Sequential Estimators **4**

4.1 Recursive Empirical Mean 4

5 Stochastic Gradient Descent **4**

5.1 Problem Formulation 4

5.2 Convex Functions 4

6 Back to the Bakery Problem **5**

1 The Bakery Problem

Note

It is interesting to start from the expression of the reward:

$$R_t = g(A_t \wedge U_t) - \ell(A_t - U_t)_+$$

$$= g(A_t) - (\ell + g)(A_t - U_t)_+$$

Since $\ell \cdot U_t \geq A_t$, we have: $g(A_t \wedge U_t) - \ell(A_t - U_t)_+ = g(A_t)$
And $g(A_t) - (l + g)(A_t - U_t)_+ = gA_t$
If $A_t \geq U_t$: $= g(U_t) - \ell(A_t - U_t) = (g + \ell)(U_t) - \ell(A_t)$

Let $\phi(a) = \mathbb{E}[R_t | A_t = a] = ga - (\ell + g)\mathbb{E}[(a - U_t)_+]$
$= ga - (g + \ell) \int_0^a \mathbb{P}(U_t \leq u) du$
$= ga - (g + \ell) \int_0^a (a - u) dF(u)$ where F is the CDF of U_t.

Continuing the calculations, we find the maximum at $a^* = F^{-1}\left(\frac{g}{g+\ell}\right)$, where F^{-1} is the quantile function of the demand.

2 Monte-Carlo Policy Evaluation

2.1 Theoretical Background

Note

Monte-Carlo Evaluation:
Let $X_1, \ldots, X_N$ be N independent random variables with the same law $\mu \in \mathcal{M}_1(\mathbb{R})$
Let $\mu = \mathbb{E}[X_1]$
Estimation of μ: $\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^N X_i$
The law of large numbers ensures that $\hat{\mu}_N \to \mu$ almost surely when $N \to +\infty$.

2.2 Variance Analysis

The variance of the estimator:

$$\mathrm{Var}(\hat{\mu}_N) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N X_i\right) = \frac{1}{N^2}\sum_{i=1}^N \mathrm{Var}(X_i) = \frac{\sigma^2}{N}$$

where $\sigma^2 = \mathrm{Var}(X_1)$

$\hat{\mu}_N$ is an unbiased estimator of μ with variance that decreases as $1/N$.

To have $|\hat{\mu}_N - \mu| \leq \epsilon$, we take $n = \frac{\mathrm{Var}(X_1)}{\epsilon^2}$, which gives:

$$\mathbb{E}[(\hat{\mu}_N - \mu)^2] = \frac{\mathrm{Var}(X_1)}{\frac{\mathrm{Var}(X_1)}{\epsilon^2}} = \epsilon^2$$

2.3 Variance Bounds

If $X_1 \in [a, b]$ almost surely, the worst case is a Bernoulli distribution and the variance is bounded by $\frac{(b-a)^2}{4}$ because:

$$\mathrm{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 \leq \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 \leq \frac{(b-a)^2}{4}$$

2.4 Central Limit Theorem

$$\sqrt{N}(\hat{\mu}_N - \mu) \to \mathcal{N}(0, \sigma^2) \text{ in distribution}$$

By normalizing: $\sqrt{N}\frac{(\hat{\mu}_N - \mu)}{\sqrt{\mathrm{Var}(X_1)}} \to \mathcal{N}(0, 1)$

This means that for N large enough:

$$\mu \in \left[\hat{\mu}_N - 2\frac{\sqrt{\mathrm{Var}(X_1)}}{\sqrt{N}}, \hat{\mu}_N + 2\frac{\sqrt{\mathrm{Var}(X_1)}}{\sqrt{N}}\right]$$

with 95% probability.

2.5 Variance Estimation

$$\hat{\sigma}_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu}_N)^2 \to \sigma^2$$

almost surely when $N \to +\infty$.

3 Non-Asymptotic Bounds

3.1 Bienaymé-Tchebychev Inequality

$$\mathbb{P}(|\hat{\mu}_N - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(X_1)}{N\epsilon^2}$$

For a 5% error, we take $N = \frac{\mathrm{Var}(X_1)}{0.05\epsilon^2}$

3.2 Hoeffding Inequality

If $\mathbb{P}(X_1 \in [a, b]) = 1$ then:

$$\mathbb{P}(|\hat{\mu}_N - \mu| \geq \epsilon) \leq 2\exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right)$$

For a 5% error, we take $N = \frac{(b-a)^2}{2\epsilon^2}\log\left(\frac{2}{0.05}\right)$

4 Sequential Estimators

4.1 Recursive Empirical Mean

$$\hat{\mu}_{t+1} = \phi(\hat{\mu}_t, X_{t+1})$$

Recursive computation of empirical mean:

$$\hat{\mu}_{t+1} = \frac{t}{t+1}\hat{\mu}_t + \frac{1}{t+1}X_{t+1}$$

By induction, we can show that $\hat{\mu}_t = \frac{1}{t}\sum_{i=1}^{t} X_i$

$$\hat{\mu}_{t+1} = \hat{\mu}_t + \frac{1}{t+1}(X_{t+1} - \hat{\mu}_t)$$

This is a stochastic approximation algorithm that resembles gradient descent.

5 Stochastic Gradient Descent

5.1 Problem Formulation

The idea is to minimize $\phi(a) = \mathbb{E}[(X_1 - a)^2]$ by differentiating:

$$\phi'(a) = 2\mathbb{E}[a - X_1]$$

Gradient Descent:

$$u_0 \in \mathbb{R} \quad (1)$$
$$u_{t+1} = u_t - \rho_t \nabla F(u_t) \quad (2)$$

Stochastic Gradient Descent:

$$u_0 \in \mathbb{R} \quad (3)$$
$$u_{t+1} = u_t - \rho_t \hat{\nabla} F(u_t) \quad (4)$$

with $\mathbb{E}[\hat{\nabla} F_t(u_t)] = \nabla F_t(u_t)$

5.2 Convex Functions

Definition

A function $F : \mathbb{R}^d \to \mathbb{R}$ is convex if:

$$\forall x, y \in \mathbb{R}^d, \forall \lambda \in [0, 1], F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y)$$

To choose the step size, we can use the Lipschitz inequality:

$$|F(x) - F(y)| \leq L\|x - y\|$$

where L is the Lipschitz constant.

Theorem

Gradient Descent Convergence
With this assumption, we can take $\rho_t = \frac{R}{L\sqrt{t}}$ and obtain:

$$F(u_T) - F(u^*) \leq \frac{RL}{\sqrt{T}}$$

We must choose R such that $\|u_0 - u^*\| \leq R$ where $u^* = \arg\min_u F(u)$.

6 Back to the Bakery Problem

When we don't know the distribution of U_t:

Note

Idea 1: Plug-in estimate of the distribution
$F^{-1}\left(\frac{g}{g+\ell}\right)$ requires F
$F(u) = \mathbb{P}(U_t \leq u)$
$\hat{F}_t(u) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{U_i \leq u}$ empirical CDF
But we don't have the past demands.
We can use estimation with censoring such as the Kaplan-Meier estimator.
$S(u) = \mathbb{P}(U_t > u)$

$$S(u) = \prod_{j=0}^{u} \mathbb{P}(U_t > j | U_t \geq j) = \prod_{j=0}^{u} (1 - \mathbb{P}(U_t = j | U_t \geq j))$$

Estimate by:

$$\hat{d}_j = \frac{\mathrm{Card}(\{i : \tilde{U}_i = j, A_i > j\})}{\mathrm{Card}(\{i : \hat{U}_i \geq j, A_i > j\})}$$