Stochastic Approximation

- Suppose X_t $(t = 1, \ldots, \infty)$ is a family if identically distributed random variables. We want to compute the mean.

- Batch algorithm:

$$\overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$$

- Incremental algorithm:

$$\overline{X} := \frac{n-1}{n} \overline{X} + \frac{1}{n} X_t$$

- Let $\alpha_t = \frac{1}{t}$, we can rewrite this as

$$\overline{X} := (1 - \alpha_t)\overline{X} + \alpha_t X_t$$

  or alternatively as

$$\overline{X} := \overline{X} + \alpha_t (X_t - \overline{X})$$

- **Robbins & Munro**. This converges to the right answer with probability 1 as long as

$$\lim_{T \to \infty} \sum_{t=1}^{T} \alpha_t = \infty \quad \text{and} \quad \lim_{T \to \infty} \sum_{t=1}^{T} \alpha_t^2 < \infty$$

A More General Stochastic Approximation Result

Consider the following case:

$$\overline{X} - (1 - \alpha_t)\overline{X} + \alpha_t(X_t + w_t)$$

where $E[w_t] = 0$.

This will also converge to the true mean.

Average-Reward Reinforcement Learning

- **Setting:** Continuing task, but no discounting
- **Goal:** Maximize the average reward per time step, ρ^π

$$\rho^\pi = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} E_\pi\{r_t\}$$

- **Assumption:** The Markov process is ergodic: Under any policy, any state is reachable from any other state.
  This means the average reward is the same no matter what state you start in.
- **Value function:** The infinite sum of *differences* between the reward received and the average reward.

$$\tilde{V}^\pi(s) = \sum_{k=1}^{\infty} E_\pi\left\{ r_{t+k} - \rho^\pi \middle| \, s_t = s \right\}$$

The R-learning algorithm

- **Basic Idea: Maintain current estimate of ρ and of $\tilde{Q}(s,a)$**

$$\tilde{Q}^{\pi} = \sum_{k=1}^{\infty} E_{\pi}\left\{ r_{t+k} - \rho^{\pi} \,|\, s_t = s, a_t = a \right\}$$

Initialize ρ and $\tilde{Q}(s,a)$, for all s, a, arbitrarily

Repeat forever:

$s :=$ current state

Choose action a using exploration policy

Take action a, observe r and s'

$\tilde{Q}(s,a) := (1-\alpha)\tilde{Q}(s,a) + \alpha[r + \max_{a'} \tilde{Q}(s',a') - \rho]$

If a was a greedy action, then update ρ:

$\rho := (1-\beta)\rho + \beta[(r + \max_{a'} \tilde{Q}(s',a')) - \max_{a} \tilde{Q}(s,a)]$