Robbins, Monro: A Stochastic Approximation Method

Robert Bassett

University of California Davis

Student Run Optimization Seminar
Oct 10, 2017

Motivation

You are a carrot farmer. The amount of carrots that you plant plays a part in how much carrots cost in the store, and hence how much profit you make on your harvest.

Motivation

You are a carrot farmer. The amount of carrots that you plant plays a part in how much carrots cost in the store, and hence how much profit you make on your harvest.

- Too many carrots planted $\to$ glut in the market and low profit.
- Too few carrots planted $\to$ unexploited demand for carrots and low profit.

Motivation

You are a carrot farmer. The amount of carrots that you plant plays a part in how much carrots cost in the store, and hence how much profit you make on your harvest.

- Too many carrots planted $\rightarrow$ glut in the market and low profit.
- Too few carrots planted $\rightarrow$ unexploited demand for carrots and low profit.

The connection between carrots planted and profit is complicated! We will model this as a stochastic optimization problem, in an attempt to deal with the random components of this model (weather, economic climate, transportation costs, etc.).

Goal: Find a carrot planting decision θ^* which minimizes your expected loss.

- Model loss as a *random* function of your planting decision θ.

$$\text{Loss} \sim q(x, \theta) \qquad \textit{known}$$

  - x: random vector of external factors. Independent of θ.

$$x \sim F(x) \qquad \textit{unknown}$$

Goal: Find a carrot planting decision θ^* which minimizes your expected loss.

- Model loss as a *random* function of your planting decision θ.

$$\text{Loss} \sim q(x, \theta) \qquad \textit{known}$$

  - x: random vector of external factors. Independent of θ.

$$x \sim F(x) \qquad \textit{unknown}$$

- Carrot Planter's Problem (CPP)

$$\min_{\theta \in \Theta} Q(\theta) = \mathbb{E}[q(x, \theta)]$$

Goal: Find a carrot planting decision θ^* which minimizes your expected loss.

- Model loss as a *random* function of your planting decision θ.

$$\text{Loss} \sim q(x, \theta) \qquad \textit{known}$$

  - x: random vector of external factors. Independent of θ.

$$x \sim F(x) \qquad \textit{unknown}$$

- Carrot Planter's Problem (CPP)

$$\min_{\theta \in \Theta} Q(\theta) = \mathbb{E}[q(x, \theta)]$$

- $q : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$
- $\Theta \subseteq \mathbb{R}^n$, set of possible decisions.

$$\min_{\theta \in \Theta} Q(\theta) = \mathbb{E}[q(x, \theta)] \qquad (CPP)$$

$$\min_{\theta \in \Theta} Q(\theta) = \mathbb{E}[q(x,\theta)] \qquad (CPP)$$

Assumptions

$$\min_{\theta \in \Theta} Q(\theta) = \mathbb{E}[q(x, \theta)] \qquad (CPP)$$

Assumptions

1. With probability 1, $q(x, \theta)$ strongly convex in θ, with uniform modulus K. That is,

$$\langle \theta_1 - \theta_2, \nabla q(x, \theta_1) - \nabla q(x, \theta_2) \rangle \geq K \left|\left|\theta_1 - \theta_2\right|\right|^2 \quad x \text{ a.s.}$$

$$\min_{\theta \in \Theta} Q(\theta) = \mathbb{E}[q(x, \theta)] \qquad (CPP)$$

Assumptions

1. With probability 1, $q(x, \theta)$ strongly convex in θ, with uniform modulus K. That is,

$$\langle \theta_1 - \theta_2, \nabla q(x, \theta_1) - \nabla q(x, \theta_2) \rangle \geq K \left|\left|\theta_1 - \theta_2\right|\right|^2 \quad x \text{ a.s.}$$

2. We have access to a random variable $y \sim P(y, \theta)$ with the property that

$$\mathbb{E}[y|\theta] = \nabla Q(\theta).$$

$$\min_{\theta \in \Theta} Q(\theta) = \mathbb{E}[q(x,\theta)] \qquad (CPP)$$

Assumptions

1. With probability 1, $q(x,\theta)$ strongly convex in θ, with uniform modulus K. That is,

$$\langle \theta_1 - \theta_2, \nabla q(x,\theta_1) - \nabla q(x,\theta_2) \rangle \geq K \, ||\theta_1 - \theta_2||^2 \quad x \text{ a.s.}$$

2. We have access to a random variable $y \sim P(y,\theta)$ with the property that

$$\mathbb{E}[y|\theta] = \nabla Q(\theta).$$

3. y has second moment uniformly bounded in θ.

$$\int_Y ||y||^2 \; dP(y,\theta) \leq C^2 \quad \forall \theta \in \Theta.$$

$$\min_{\theta \in \Theta} Q(\theta) = \mathbb{E}[q(x,\theta)] \qquad (CPP)$$

Assumptions

1. With probability 1, $q(x,\theta)$ strongly convex in θ, with uniform modulus K. That is,

$$\langle \theta_1 - \theta_2, \nabla q(x,\theta_1) - \nabla q(x,\theta_2) \rangle \geq K \left|\left|\theta_1 - \theta_2\right|\right|^2 \quad x \text{ a.s.}$$

2. We have access to a random variable $y \sim P(y,\theta)$ with the property that

$$\mathbb{E}[y|\theta] = \nabla Q(\theta).$$

3. y has second moment uniformly bounded in θ.

$$\int_Y \left|\left|y\right|\right|^2 \, dP(y,\theta) \leq C^2 \quad \forall \theta \in \Theta.$$

4. Q is differentiable, so (CPP) is equivalent to solving $\nabla Q(\theta) = 0$.

(2) says we must have access to an unbiased estimator of our gradient.

(2) says we must have access to an unbiased estimator of our gradient.
For many common loss functions this is true, e.g. squared loss:

$$
\begin{aligned}
q(x,\theta) &= \frac{1}{2}\mathbb{E}[||A\theta - x||^2] \\
\Rightarrow y \sim \nabla q(x,\theta) &= A^T(A\theta - x) \text{ satisfies} \\
\mathbb{E}[y|\theta] &= \mathbb{E}[\nabla q(x,\theta)|\theta] = \mathbb{E}[A^T(A\theta - x)|\theta] = \nabla Q(\theta)
\end{aligned}
$$

The last equality follows from the mild regularity assumptions that allow us to interchange derivative with integral. Know them!

- We would like to find a minimizer of $Q(\theta)$ without ever attempting to calculate the expectation that defines it.
- Why? It involves random factors that are nasty!
- Instead, we will build a sequence θ_n which depends on random samples from $y \sim P(y, \theta)$ for different θ.

- We would like to find a minimizer of $Q(\theta)$ without ever attempting to calculate the expectation that defines it.
- Why? It involves random factors that are nasty!
- Instead, we will build a sequence θ_n which depends on random samples from $y \sim P(y, \theta)$ for different θ.
- Hope to generate iterates from *samples* which allow us to minimize the expectation.
- We want θ_n to be a *consistent* estimator of θ^*. That is

$$\forall \epsilon \quad \exists N \quad \text{s.t.} \quad n \geq N \Rightarrow P(||\theta_n - \theta^*||_2 > \epsilon) < \epsilon$$

i.e. $\theta_n \to \theta$ in probability.

Facts of Life Give references!

1. L^2-convergence implies convergence in probability. That is,

$$\mathbb{E}[||\theta_n - \theta||_2] \to 0 \Rightarrow ||\theta_n - \theta||_2 \to 0 \text{ in probability.}$$

Facts of Life Give references!

1. L^2-convergence implies convergence in probability. That is,

$$\mathbb{E}[||\theta_n - \theta||_2] \to 0 \Rightarrow ||\theta_n - \theta||_2 \to 0 \text{ in probability.}$$

2. A convex function defines a monotone operator. That is, if $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$$

Facts of Life Give references!

1. L^2-convergence implies convergence in probability. That is,

$$\mathbb{E}[||\theta_n - \theta||_2] \to 0 \Rightarrow ||\theta_n - \theta||_2 \to 0 \text{ in probability.}$$

2. A convex function defines a monotone operator. That is, if $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$$

3. $f(\xi, x)$ convex in x for some ξ almost everywhere implies

$$F(x) = \mathbb{E}[f(\xi, x)]$$

is convex.

Theorem

Let a_n be a positive sequence which satisfies:

- $$\sum_{n=0}^{\infty} a_n^2 < \infty$$
- $$\sum_{n=0}^{\infty} \frac{a_n}{a_0 + \ldots + a_{n-1}} = \infty$$

For some initial θ_0, define the sequence

$$\theta_{n+1} = \theta_n - a_n y_n$$

Where $y_n \sim P(y|\theta_n)$. Then $\theta_n \to \theta^*$ in probability.

Proof:

Proof:
Because of F.O.L. 1, we will instead show L^2 convergence.

Proof:

Proof:
Define $b_n = \mathbb{E}[||\theta_n - \theta^*||^2]$. We want to show that $b_n \to 0$.

Proof:
Define $b_n = \mathbb{E}[||\theta_n - \theta^*||^2]$. We want to show that $b_n \to 0$.
Have that:

$$b_{n+1} = \mathbb{E}[||\theta_{n+1} - \theta^*||^2]$$

Proof:
Define $b_n = \mathbb{E}[||\theta_n - \theta^*||^2]$. We want to show that $b_n \to 0$.
Have that:

$$b_{n+1} = \mathbb{E}[||\theta_{n+1} - \theta^*||^2]$$

$$= \mathbb{E}[\mathbb{E}[||\theta_{n+1} - \theta^*||^2 \, |\theta_n]]$$

Proof:
Define $b_n = \mathbb{E}[||\theta_n - \theta^*||^2]$. We want to show that $b_n \to 0$.
Have that:

$$b_{n+1} = \mathbb{E}[||\theta_{n+1} - \theta^*||^2]$$

$$= \mathbb{E}[\mathbb{E}[||\theta_{n+1} - \theta^*||^2 \, |\theta_n]]$$

$$= \mathbb{E}\left[\int_Y ||(\theta_n - \theta^*) - a_n y||^2 \, dP(y|\theta_n)\right]$$

Proof:
Define $b_n = \mathbb{E}[||\theta_n - \theta^*||^2]$. We want to show that $b_n \to 0$.
Have that:

$$b_{n+1} = \mathbb{E}[||\theta_{n+1} - \theta^*||^2]$$

$$= \mathbb{E}[\mathbb{E}[||\theta_{n+1} - \theta^*||^2 \, |\theta_n]]$$

$$= \mathbb{E}\left[\int_Y ||(\theta_n - \theta^*) - a_n y||^2 \; dP(y|\theta_n)\right]$$

$$= b_n + a_n^2 \mathbb{E}\left[\int_Y ||y||^2 \; dP(y|\theta_n)\right] - 2a_n \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n)\rangle]$$

$$= b_n + a_n^2 \underbrace{\mathbb{E}\left[\int_Y ||y||^2 \, dP(y|\theta_n)\right]} -2a_n \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n)\rangle]$$

$$= b_n + a_n^2 \underbrace{\mathbb{E}\left[\int_Y ||y||^2 \; dP(y|\theta_n)\right]} - 2a_n \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]$$

$$\mathbb{E}\left[\int_Y ||y||^2 \; dP(y|\theta_n)\right] < C^2$$

by Assumption 3

$$\leq b_n + a_n^2 C^2 - 2a_n \underbrace{\mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]}$$

$$\leq b_n + a_n^2 C^2 - 2a_n \underbrace{\mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]}$$

Because $\theta^* \in \operatorname{argmin}_\theta Q(\theta)$, $\nabla Q(\theta^*) = 0$.

$$\leq b_n + a_n^2 C^2 - 2a_n \underbrace{\mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]}$$

Because $\theta^* \in \mathrm{argmin}_\theta\, Q(\theta)$, $\nabla Q(\theta^*) = 0$.

$$\mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle] = \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) - \nabla Q(\theta^*) \rangle]$$

$$\leq b_n + a_n^2 C^2 - 2a_n \underbrace{\mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]}$$

Because $\theta^* \in \text{argmin}_\theta\, Q(\theta)$, $\nabla Q(\theta^*) = 0$.

$$\mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle] = \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) - \nabla Q(\theta^*) \rangle]$$

$$\geq 0$$

by F.O.L. 2 and 3

$$0 \leq b_{n+1} \leq b_n + a_n^2 C^2 - 2a_n \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]$$

$$0 \leq b_{n+1} \leq b_n + a_n^2 C^2 - 2a_n \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]$$

$$\Rightarrow$$

$$0 \leq b_{n+1} \leq b_0 + C^2 \sum_{i=0}^{n} a_n^2 - 2 \sum_{i=0}^{n} a_i \mathbb{E}[\langle \theta_i - \theta^*, \nabla Q(\theta_i) \rangle]$$

$$0 \leq b_{n+1} \leq b_n + a_n^2 C^2 - 2a_n \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]$$

$$\Rightarrow$$

$$0 \leq b_{n+1} \leq b_0 + C^2 \sum_{i=0}^{n} a_n^2 - 2 \sum_{i=0}^{n} a_i \mathbb{E}[\langle \theta_i - \theta^*, \nabla Q(\theta_i) \rangle]$$

$$\Rightarrow$$

$$0 \leq \sum_{i=0}^{n} a_i \mathbb{E}[\langle \theta_i - \theta^*, \nabla Q(\theta_i) \rangle] \leq \frac{1}{2} \left(b_0 + C^2 \sum_{i=1}^{n} a_i^2 \right)$$

$$0 \leq b_{n+1} \leq b_n + a_n^2 C^2 - 2a_n \mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle]$$

$$\Rightarrow$$

$$0 \leq b_{n+1} \leq b_0 + C^2 \sum_{i=0}^{n} a_n^2 - 2 \sum_{i=0}^{n} a_i \mathbb{E}[\langle \theta_i - \theta^*, \nabla Q(\theta_i) \rangle]$$

$$\Rightarrow$$

$$0 \leq \sum_{i=0}^{n} a_i \mathbb{E}[\langle \theta_i - \theta^*, \nabla Q(\theta_i) \rangle] \leq \frac{1}{2} \left(b_0 + C^2 \sum_{i=1}^{n} a_i^2 \right)$$

Taking the limit as $n \to \infty$, gives

$$0 \leq \lim_{n \to \infty} b_n < \infty$$

Great! This shows that $\lim_{n\to\infty} b_n$ exists. But is it equal to 0?

Great! This shows that $\lim_{n\to\infty} b_n$ exists. But is it equal to 0? Consider the sequence

$$k_n = \frac{K}{a_1 + ... + a_{n-1}}$$

Great! This shows that $\lim_{n\to\infty} b_n$ exists. But is it equal to 0?
Consider the sequence

$$k_n = \frac{K}{a_1 + ... + a_{n-1}}$$

We have that

$$\mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n)\rangle] \geq K\mathbb{E}[||\theta_n - \theta^*||^2]$$

By strong convexity.

Great! This shows that $\lim_{n\to\infty} b_n$ exists. But is it equal to 0?
Consider the sequence

$$k_n = \frac{K}{a_1 + ... + a_{n-1}}$$

We have that

$$\mathbb{E}[\langle \theta_n - \theta^*, \nabla Q(\theta_n) \rangle] \geq K \mathbb{E}[||\theta_n - \theta^*||^2]$$

By strong convexity.
Rewriting, we then have

$$\geq K b_n \geq k_n b_n$$

for large enough n.

Great! This shows that $\lim_{n\to\infty} b_n$ exists. But is it equal to 0?
Consider the sequence

$$k_n = \frac{K}{a_1 + ... + a_{n-1}}$$

We have that

$$\mathbb{E}[\langle\theta_n - \theta^*, \nabla Q(\theta_n)\rangle] \geq K\mathbb{E}[||\theta_n - \theta^*||^2]$$

By strong convexity.
Rewriting, we then have

$$\geq Kb_n \geq k_n b_n$$

for large enough n.
We proved previously that summing the a_n times inner product above gives a convergent sequence! This gives that, since k_n and b_n are both positive,

$$\sum_n^{\infty} a_n k_n b_n < \infty.$$

By assumption on a_n, we have also that

$$a_n k_n = \frac{K a_n}{a_1 + ... + a_{n-1}} \to \infty$$

By assumption on a_n, we have also that

$$a_n k_n = \frac{K a_n}{a_1 + ... + a_{n-1}} \to \infty$$

So $\sum a_n k_n \to \infty$.

We have established

We have established

1. $$\sum_{i=1}^{\infty} a_i k_i b_i < \infty$$

2. $$\sum_{i=1}^{\infty} a_i k_i \to \infty$$

We have established

1. $$\sum_{i=1}^{\infty} a_i k_i b_i < \infty$$

2. $$\sum_{i=1}^{\infty} a_i k_i \to \infty$$

So...

We have established

1. $$\sum_{i=1}^{\infty} a_i k_i b_i < \infty$$

2. $$\sum_{i=1}^{\infty} a_i k_i \to \infty$$

So...

$$b_n \to 0!$$

We have established

1. $$\sum_{i=1}^{\infty} a_i k_i b_i < \infty$$

2. $$\sum_{i=1}^{\infty} a_i k_i \to \infty$$

So...

$$b_n \to 0!$$

$\square$

So are there any sequences that satisfy the conditions in the theorem?

So are there any sequences that satisfy the conditions in the theorem?

- $$\sum_{n=0}^{\infty} a_n^2 < \infty$$
- $$\sum_{n=0}^{\infty} \frac{a_n}{a_1 + ... + a_{n-1}} = \infty$$

So are there any sequences that satisfy the conditions in the theorem?

- $$\sum_{n=0}^{\infty} a_n^2 < \infty$$
- $$\sum_{n=0}^{\infty} \frac{a_n}{a_1 + \ldots + a_{n-1}} = \infty$$

Take $a_n = \frac{1}{n}$. Square summable and

$$\sum_{n=0}^{\infty} \frac{\frac{1}{n}}{\frac{1}{1} + \ldots + \frac{1}{n-1}} \approx \sum_{n=0}^{\infty} \frac{1}{n \ln(n-1)} \geq \sum_{n=0}^{\infty} \frac{1}{n \ln n} \to \infty$$

Extensions and loose ends

- Actually converges with probability 1 (Blum).
- Convergence with rates
  - $\mathbb{E}[Q(\theta_n) - Q(\theta^*)] \in O(n^{-1})$ (with strong convexity)
  - $\mathbb{E}[Q(\theta_n) - Q(\theta^*)] \in O(n^{-\frac{1}{2}})$ (without strong convexity)
- $\frac{\theta_n - \theta^*}{\sqrt{n}}$ is asymptotically normal. (Sacks)
- $\sqrt{n}$ rate *cannot* be beat for general convex case. (Nemirovski et al)