Convex Optimisation

Matthew Tam
(University of Melbourne)
matthew.tam@unimelb.edu.au

ACE Network
S1 2021

---

* Assignment 4
  - due June 3rd @ 5pm

Continued from last time

Recall that

$$D = \{(x_1, \dots, x_m) \in \underbrace{\mathbb{R}^n \times \dots \times \mathbb{R}^n}_{\equiv \mathbb{R}^{nm}} : x_1 = \dots = x_m\}$$

Last time, we needed to compute

$$\text{prox}_{i_D} = P_D.$$

Proposition 3.5.4

Let $\mathbf{x} = (x_1, \dots, x_m) \in \mathbb{R}^n \times \dots \times \mathbb{R}^n$. Then

$\mathbf{p} = (p, \dots, p) = P_D(\mathbf{x})$ where

$$p = \frac{1}{m} \sum_{j=1}^{m} x_j.$$

Proof.

We prove the result by using Corollary 2.3.6. We first note that D is nonempty, closed and convex. Next note that $\mathbf{p} \in D$ by definition. Finally, let $\mathbf{c} = (c, \ldots, c) \in D$ and observe

$$\begin{aligned}
&\langle \mathbf{x} - \mathbf{p}, \mathbf{c} - \mathbf{p} \rangle \\
&= \sum_{i=1}^{m} \langle x_i - p, c - p \rangle \\
&= \sum_{i=1}^{m} (x_i, c - p) - m \langle p, c - p \rangle \\
&= \sum_{i=1}^{m} \langle x_i, c - p \rangle - m \left\langle \frac{1}{m} \sum_{j=1}^{m} x_j, c - p \right\rangle \\
&= \sum_{i=1}^{m} (x_i, c - p) - \sum_{i=1}^{m} (x_i, c - p) \\
&= 0,
\end{aligned}$$

which completes the proof. ∎

3.6 Stochastic Gradient Descent

Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{where} \quad f(x) := \frac{1}{m} \sum_{i=1}^{m} f_i(x)$$

when m is large and $f_i : \mathbb{R}^n \to \mathbb{R}$ are convex and differentiable. Under these assumptions, f is also convex and differentiable, so we could apply gradient descent (Section 3.2). That is,

$$x_{k+1} = x_k - \lambda \underbrace{\nabla f(x_k)}_{\text{approximate this.}} = x_k - \frac{\lambda}{m} \sum_{i=1}^{m} \nabla f_i(x_k).$$

If m is large, the sum on the RHS might take too long to compute.

We require the stepsize sequence (λ_k) to satisfy

$$\sum_{i=0}^{\infty} \lambda_k^2 < +\infty, \quad \text{and} \qquad \left[\lambda_k = \frac{1}{k+1}\right]$$

$$\sum_{i=0}^{\infty} \lambda_k = +\infty.$$

Given an initial point $x_0 \in \mathbb{R}^n$, generate sequences (i_k) and (x_k) according to

$$\begin{cases} i_k \sim \text{unif}\{1, \dots, m\} \\ x_{k+1} = x_k - \lambda_k \nabla f_{i_k}(x_k) \end{cases} \qquad \forall k \in \mathbb{N}. \tag{3.27}$$

Here the notation '$i_k \sim \text{unif}\{1, \dots, m\}$' means in the kth iteration, the integer i_k is choosen at random from $\{1, \dots, m\}$.

with outcome having equal probability.

Definition 3.6.1

Let Z be a random variable with a finite number of outcomes $z_1, z_2, \dots, z_m$ occurring with probabilities $p_1, p_2, \dots, p_m$. The expectation of Z is defined as

$$\mu := \mathbb{E}[Z] := \sum_{i=1}^{m} p_i z_i .$$

Note that expectation is linear: if Z_1 and Z_2 are random variables and $a \in \mathbb{R}$, then

$$\mathbb{E}[Z_1 + a Z_2] = \mathbb{E}[Z_1] + a\,\mathbb{E}[Z_2].$$

Now, since $i_k \sim \text{unif}\{1, \dots, m\}$ and x_k doesn't depend on i_k, we have.

$$\mathbb{E}\left[f_{i_k}(x_k)\right] = \sum_{i=1}^{m} \frac{1}{m} \cdot f_i(x_k) = \frac{1}{m} \sum_{i=1}^{m} f_i(x_k)$$

$$= f(x_k)$$

Soln.

$$\mathbb{E}\left[f_{i_k}(x^*)\right] = \frac{1}{m} \sum_{i=1}^{m} f_i(x^*) = f(x^*)$$

$$\mathbb{E}\left[\nabla f_{i_k}(x_k)\right] = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x_k) = \nabla f(x_k)$$

$\vdots$

Theorem 3.6.2

Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable for all $i \in \{1, \dots, m\}$ and set $f(x) := \frac{1}{m} \sum_{i=1}^{m} f_i(x)$. Let $x^* \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$. Given $x_0 \in \mathbb{R}^n$, let (x_k) be the sequence generated by (3.27). Further suppose $\exists \sigma > 0$ such that $\|\nabla f_i(x)\| \leq \sigma$ for all $x \in \mathbb{R}^n$ and all $i \in \{1, \dots, m\}$. Then

$\lambda_i \to 0$ as $i \to \infty$

$$\mathbb{E}[f(\bar{x}_k)] - f(x^*) \leq \frac{\|x_0 - x^*\|^2 + \sigma^2 \sum_{j=0}^{k} \lambda_j^2}{2\left(\sum_{j=0}^{k} \lambda_j\right)} \to 0 \quad \forall k \in \mathbb{N}.$$

$\sum_{j=0}^{k} \lambda_j \to +\infty$ as $k \to \infty$

where $\bar{x}_k := \sum_{j=0}^{k} \left(\frac{\lambda_j}{\sum_{i=0}^{k} \lambda_i} x_j \right)$.

In particular, $\mathbb{E}[f(\bar{x}_k)] \to f(x^*)$ as $k \to \infty$.

Proof

Since f_i is convex and differentiable, Proposition 2.2.3(a) implies

$$f_i(x^*) \geq f_i(x_k) + \langle \nabla f_i(x_k), x^* - x_k \rangle.$$

Expand the norm squared gives

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 \\
&= \|(x_k - \lambda_k \nabla f_{i_k}(x_k)) - x^*\|^2 \\
&= \|x_k - x^*\|^2 - 2\lambda_k \langle \nabla f_{i_k}(x_k), x_k - x^* \rangle \\
&\qquad + \lambda_k^2 \|\nabla f_{i_k}(x_k)\|^2
\end{aligned}$$

$$\leq \|x_k - x^*\|^2 + 2\lambda_k \langle \nabla f_{i_k}(x_k), x^* - x_k \rangle + \lambda_k^2 \sigma^2 .$$

Taking the expectation of both sides, we obtain

$$\mathbb{E}\left\{ \|x_{k+1} - x^*\|^2 \right\}$$

$$\leq \mathbb{E}\left\{ \|x_k - x^*\|^2 + 2\lambda_k \langle \nabla f_{i_k}(x_k), x^* - x_k \rangle + \lambda_k^2 \sigma^2 \right\}$$

$$= \|x_k - x^*\|^2 + 2\lambda_k \mathbb{E}\left[\langle \nabla f_{i_k}(x_k), x^* - x_k \rangle \right] + \lambda_k^2 \sigma^2$$

$$= \|x_k - x^*\|^2 + 2\lambda_k \langle \mathbb{E}[\nabla f_{i_k}(x_k)], x^* - x_k \rangle + \lambda_k^2 \sigma^2$$

$$= \|x_k - x^*\|^2 + 2\lambda_k \langle \nabla f(x_k), x^* - x_k \rangle + \lambda_k^2 \sigma^2 .$$

$$\leq \|x_k - x^*\|^2 + 2\lambda_k (f(x^*) - f(x_k)) + \lambda_k^2 \sigma^2 .$$

Rearranging this inequality gives

$$2\lambda_k (f(x_k) - f(x^*)) \leq \|x_k - x^*\|^2 - \mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] + \lambda_k^2 \sigma^2 .$$

Now, consider x_k as being determined by the random variable $i_1, i_2, \dots i_{k-1}$. Taking the expectation of both sides gives

$$2\lambda_k \left(\mathbb{E}[f(x_k)] - f(x^*) \right)$$

$$\leq \mathbb{E}\left[\|x_k - x^*\|^2 \right] - \mathbb{E}\left[\|x_{k+1} - x^*\|^2 \right] + \lambda_k^2 \sigma^2 .$$

Summing this inequality for $j = 0, \dots, k$ gives

$$2 \sum_{j=0}^{k} \lambda_j \left(\mathbb{E}[f(x_j)] - f(x^*) \right)$$

$$\leq \sum_{j=0}^{k} \left(\mathbb{E}\left[\|x_j - x^*\|^2\right] - \mathbb{E}\left[\|x_{j+1} - x^*\|^2\right] \right) + \sigma^2 \sum_{j=0}^{k} \lambda_j^2$$

$$= \mathbb{E}\left[\|x_0 - x^*\|^2\right] - \mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] + \sigma^2 \sum_{j=0}^{k} \lambda_j^2$$

$$\leq \|x_0 - x^*\|^2 + \sigma^2 \sum_{j=0}^{k} \lambda_j^2 .$$

Dividing both sides by $2\sum_{j=0}^{k} \lambda_j$ gives

$$\sum_{j=0}^{k} \left(\frac{\lambda_j}{\sum_{i=0}^{k} \lambda_i} \mathbb{E}[f(x_j)] \right) - f(x^*) \leq \frac{\|x_0 - x^*\|^2 + \sigma^2 \sum_{j=0}^{k} \lambda_j^2}{2 \sum_{j=0}^{k} \lambda_j}$$

RHS of what we want

Finally, using convexity of f, we deduce that

$$\mathbb{E}[f(\bar{x}_k)] = \mathbb{E}\left[f\left(\sum_{j=0}^{k} \left(\frac{\lambda_j}{\sum_{i=0}^{k} \lambda_i} \right) x_j \right) \right]$$

$= \bar{x}_k$ by defn.

Same.

$$\leq \mathbb{E}\left[\sum_{j=0}^{k} \left(\frac{\lambda_j}{\sum_{i=0}^{k} \lambda_i} \right) f(x_j) \right]$$

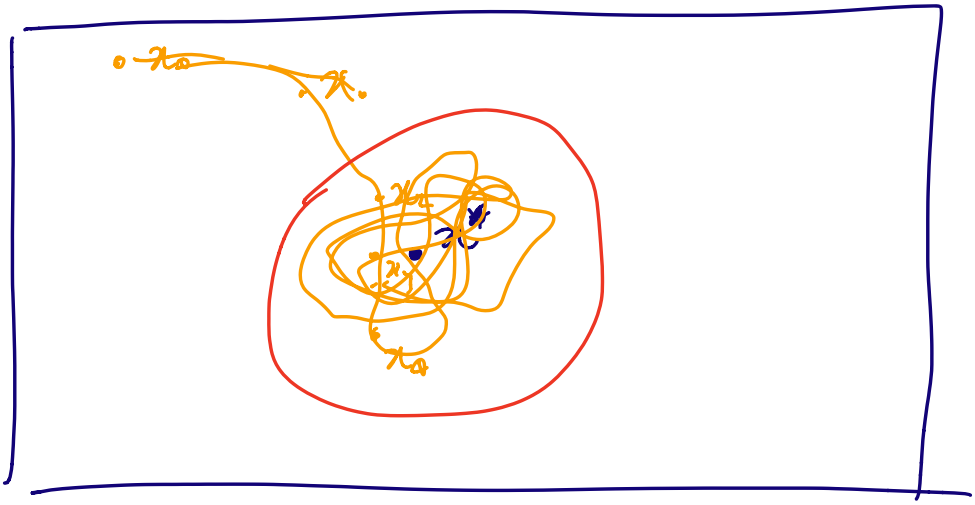
$$= \sum_{j=0}^{k} \left(\frac{\lambda_j}{\sum_{i=0}^{k} \lambda_i} \right) \mathbb{E}\left[f(x_j) \right].$$

The result follows by combining the last two inequalities. ∎

$$\begin{cases} i_k \sim \text{unif}\{1, \dots, n\} \\ x_{k+1} = x_k - \lambda \nabla f_{i_k}(x_k) \end{cases}$$

for some $\lambda > 0$. There exist $\varepsilon \geq 0$ such that

$$f(x_k) - f(x^*) \leq \varepsilon \quad \forall k \in \mathbb{N}.$$



- variance reduction techniques.