Exercise List: Sampling, Mini-batching and momenutum

Robert M. Gower

November 18, 2019

1 Introduction and definitions

Consider the problem

$$w^* \in \arg\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f_i(w) =: f(w), \tag{1}$$

where $f(w)$ is convex and differentiable.

We define a sampling vector

Definition 1.1. We say that a random vector $v \in \mathbb{R}^n$ drawn from some distribution $\mathcal{D}$ is a *sampling vector* if its mean is the vector of all ones:

$$\mathbb{E}_{\mathcal{D}}\left[v_i\right] = 1, \quad \forall i \in [n]. \tag{2}$$

With this definition we can re-write our original problem as as follows

$$\min_{x\in\mathbb{R}^d} \mathbb{E}_{\mathcal{D}}\left[f_v(w) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n v_i f_i(w)\right]. \tag{3}$$

Before we give examples of v, let us first establish some random set terminology. Let $C \subseteq [n]$ and let $e_C \overset{\text{def}}{=} \sum_{i\in C} e_i$, where $\{e_1,\ldots,e_n\}$ are the standard basis vectors in $\mathbb{R}^n$. These subsets will be selected using a random set valued map S which is known as a sampling. A sampling is uniquely characterized by choosing subset probabilities $p_C \geq 0$ for all subsets C of $[n]$:

$$\mathbb{P}\left[S = C\right] = p_C, \quad \forall C \subset [n], \tag{4}$$

where $\sum_{C\subseteq[n]} p_C = 1$.

2 Sampling

In the following exercises let $S \subset \{1,\ldots,n\}$ be a random set and let $\mathbf{1}_{i\in S}$ be the indicator function, that is

$$\mathbf{1}_{i\in S} = \begin{cases} 1 & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Ex. 1 — Let v be a sampling vector. Show that by sampling $v \sim \mathcal{D}$ the stochastic gradient $\nabla f_v(w)$ is an unbiased estimate of the full gradient with

$$\mathbb{E}\left[\nabla f_v(w)\right] = \frac{1}{n}\sum_{i=1}^{n} f_i(w) = \nabla f(w).$$

Ex. 2 — Let $\mathbb{P}\left[S = \{j\}\right] = \frac{1}{n}$ for $j = 1, \ldots, n$. Show that the random vector $v \in \mathbb{R}^n$

$$v_i = \begin{cases} n & i \in S, \\ 0 & i \notin S, \end{cases}$$

is a sampling vector. We refer to this as the *Single Element Sampling.* Furthermmore, show that

$$\nabla f_v(w) = \nabla f_j(w)$$

with probability $\frac{1}{n}$ for $j = 1, \ldots, n$.

Ex. 3 — Let $b \in \mathbb{N}$ elements and let $|S| = b$, such that every subset has equal chance of being selected. That is, given $B \subset \{1, \ldots, n\}$ with $|B| = b$ we have that

$$\mathbb{P}\left[S = B\right] \;=\; \frac{1}{\binom{n}{b}} \;=:\; \frac{1}{\frac{n!}{b!(n-b)!}}.$$

Show that $\mathbb{P}\left[i \in S\right] = \frac{b}{n}$ for $i = 1, \ldots, n$. Furthermore, show that the random vector $v \in \mathbb{R}^n$

$$v_i = \begin{cases} \frac{n}{b} & i \in S, \\ 0 & i \notin S, \end{cases}$$

is a sampling vector. We refer to this as the *b–nice Sampling.*

Ex. 4 — Let $p_i = \mathbb{P}\left[i \in S\right] > 0$ for $i = 1, \ldots, n$. That is, all elements have a non-zero probability of being sampled. Let $\hat{\mathbf{P}} = \mathrm{Diag}(p_1, ..., p_n) \in \mathbb{R}^{n \times n}$. Show that the random vector v given by

$$v \;=\; \hat{\mathbf{P}}^{-1} e_S \;=\; \sum_{i \in S} \frac{e_i}{p_i}. \tag{7}$$

is a sampling vector. We refer to this as an *arbitrary sampling.* Show that all the previous samplings are specials cases of this one.

3 Expected Smoothness

For the next exercises we need the following expected smoothness assumption and the definition of gradient noise introduced in [2, 3, 1]

Assumption 3.1 (Expected Smoothness)**.** We say that f is $\mathcal{L}$–smooth in expectation with respect to a distribution $\mathcal{D}$ if there exists $\mathcal{L} = \mathcal{L}(f, \mathcal{D}) > 0$ such that

$$\mathbb{E}_v\left[\|\nabla f_v(w) - \nabla f_v(w^*)\|^2\right] \leq 2\mathcal{L}(f(w) - f(w^*)), \tag{8}$$

for all $x \in \mathbb{R}^d$. For simplicity, we will write $(f, \mathcal{D}) \sim ES(\mathcal{L})$ to say that (8) holds. When $\mathcal{D}$ is clear from the context, we will often ignore mentioning it, and simply state that the expected smoothness constant is $\mathcal{L}$.

Definition 3.2 (Finite Gradient Noise)**.** The *gradient noise* $\sigma = \sigma(f, \mathcal{D})$, defined as follows

$$\sigma^2 \overset{\text{def}}{=} \mathbb{E}_v \left[\|\nabla f_v(w^*)\|^2 \right]. \tag{9}$$

Ex. 5 — If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, show that

$$\mathbb{E}_{\mathcal{D}} \left[\|\nabla f_v(w)\|^2 \right] \leq 4\mathcal{L}(f(w) - f(w^*)) + 2\sigma^2. \tag{10}$$

Consider the gradient noise and the samplings defined in the exercises in Section 2.

Ex. 6 — For single element sampling with $\mathbb{P}\left[v = ne_i\right] = \frac{1}{n}$ for $i = 1, \ldots, n$, show that

$$\sigma^2 = \frac{1}{n} \sum_{i \in [n]} \|\nabla f_i(w^*)\|^2. \tag{11}$$

Ex. 7 — For single element sampling with $\mathbb{P}\left[v = \frac{e_j}{p_j}\right] = p_i$ for $i = 1, \ldots, n$, show that

$$\sigma^2 = \frac{1}{n^2} \sum_{i \in [n]} \frac{1}{p_i} \|\nabla f_i(w^*)\|^2. \tag{12}$$

Ex. 8 — Given that (1) is a convex unconstrained optimization problem we have that $\nabla f(w^*) = 0$. Show that

$$\frac{1}{n^2} \sum_{i,j=1}^{n} \langle \nabla f_i(w^*), \nabla f_j(w^*) \rangle = 0.$$

Ex. 9 — **Level hard**: For b-nice sampling S with $\mathbb{P}\left[v_i = \frac{n}{b}\mathbf{1}_{i \in S}\right] = \frac{b}{n}$ show that

$$\sigma^2 = \frac{1}{nb} \cdot \frac{n-b}{n-1} \sum_{i \in [n]} \|\nabla f_i(w^*)\|^2. \tag{14}$$

[**Expected Smoothness**] Suppose that f_i is L_i–smooth and convex and consequently f is L–smooth and convex. It follows from equation (2.1.7) in Theorem 2.1.5 in [4] that

$$\|\nabla f_i(w) - \nabla f_i(y)\|^2 \leq 2L_i(f_i(w) - f_i(y) - \langle \nabla f_i(y), x - y \rangle). \tag{15}$$

Since f is L-smooth, we have

$$\|\nabla f(w) - \nabla f(y)\|^2 \leq 2L(f(w) - f(y) - \langle \nabla f(y), x - y \rangle). \tag{16}$$

For the next exercises, we will assume that (15) and (16) hold.

Ex. 10 — Show that if $\mathbb{P}\left[v = ne_i\right] = \frac{1}{n}$ then Assumption 3.1 holds with $\mathcal{L} = L_{\max}$.

Ex. 11 — **Level hard**: For b-nice sampling S with $\mathbb{P}\left[v_i = \frac{n}{b}\mathbf{1}_{i\in S}\right] = \frac{b}{n}$ show that Assumption 3.1 holds with

$$\mathcal{L} \quad = \quad \frac{n(b-1)}{b(n-1)}L + \frac{1}{b}\frac{n-b}{n-1}L_{\max}. \tag{17}$$

This formula was only recently introduced in [3] and has enabled the calculation of better mini-batch sizes in stochastic gradient methods. Note that this expected smoothness constant (17) interpolates perfectly between L and $L_{\max}$ in the sense that $\mathcal{L} = L_{\max}$ when $b = 1$ and $\mathcal{L} = L$ when $b = n$.

4 The Heavy ball/Momentum method

Ex. 12 — **Level hard**: Let $m^0 = 0 = w^0 \in \mathbb{R}^d$. Consider the Heavy Ball method give by

$$w^{t+1} = w^t - \gamma\,\nabla f(w^t) + \beta(w^t - w^{t-1}), \quad \text{for } t = 1, \ldots, T.$$

Let f be L–smooth and μ–convex and thus

$$\mu\, I \quad \preceq \quad \nabla^2 f(w) \quad \preceq L\, I, \quad \forall w \in \mathbb{R}^d. \tag{18}$$

Let $\kappa \overset{\text{def}}{=} \frac{L}{\mu}$. Let $\gamma = \dfrac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. Finally let

$$A \overset{\text{def}}{=} (1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w + s(w^* - w))ds,$$

and let $\|A\| = \max_{i=1,\ldots,d} |\lambda_i(A)|$ denote the induced norm. Show that

$$\left\| \begin{bmatrix} A & -I\beta \\ I & 0 \end{bmatrix} \right\| \quad = \quad \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

Conclude that the Heavy ball method converges at a rate of $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.

References

[1] N. Gazagnadou, R. M. Gower, and J. Salmon. "Optimal mini-batch and step sizes for SAGA". In: *ICML* (2019).

[2] R. M. Gower, P. Richtárik, and F. Bach. "Stochastic Quasi-Gradient Methods: Variance Reduction via Jacobian Sketching". In: *arxiv:1805.02632* (2018).

[3] R. M. Gower et al. "SGD: general analysis and improved rates". In: *ICML* (2019).

[4] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* 1st ed. Springer Publishing Company, Incorporated, 2014.