CS 331: Stochastic Gradient Descent Methods

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Part 8: SGD with Minibatch Sampling

Based on:

[3] R.M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P.R.
SGD: General Analysis and Improved Rates, ICML 2019

Introduction

We will describe **three families of SGD methods** which allow us to use multiple functions f_i, chosen at random, in the formation of the stochastic gradient. If we assume that each function f_i corresponds to a single training data point only, then this means that **a random collection of all training data points (a "minibatch") is processed in each iteration.**

- SGD-NS in each iteration samples & processes a single training data point only
- GD in each iteration samples & processes all the training data points
- In some sense, these methods we will **interpolate between SGD-NS (or a variant thereof) and GD.** The methods are also known as **minibatch** SGD methods.

To do define and analyze each method, we need to:

- Describe the **unbiased** stochastic gradient estimator g
- Compute **expected smoothness** constant $A'' \geq 0$ such that

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y)$$

Sampling without Replacement
(Nice Sampling)

Sampling without Replacement: Nice Sampling

Fix a minibatch size $\tau \in \{1, 2, \ldots, n\}$ and let S be a random subset of $\{1, 2, \ldots, n\}$ of size τ, chosen uniformly at random.[9] Define the gradient estimator via

$$g(x) \overset{\text{def}}{=} \frac{1}{\tau} \sum_{i \in S} \nabla f_i(x). \tag{77}$$

This estimator leads to the following SGD algorithm:

Algorithm 7 `SGD-NICE`

Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, minibatch size $\tau \in \{1, 2, \ldots, n\}$
for $k = 0, 1, 2, \ldots$ **do**
  Sample $S^k \subseteq \{1, \ldots, n\}$ uniformly from all subsets of cardinality τ
  $g^k = \frac{1}{\tau} \sum_{i \in S^k} \nabla f_i(x^k)$ obtain a stochastic gradient
  $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

[9] That is, we choose a single subset from the $\binom{n}{\tau}$ subsets of $\{1, 2, \ldots, n\}$ of cardinality τ, each with probability $1/\binom{n}{\tau}$. Such a random set S is also known in the literature under the name τ-nice sampling [?].

SGD-NICE: Unbiasedness and Expected Smoothness

Lemma 45

The gradient estimator g defined in (77) is unbiased. If we further assume that $n \geq 2$, f_i is convex and L_i-smooth for all i, and f is L-smooth, then

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \leq 2A'' D_f(x, y),$$

where

$$A'' = \frac{n - \tau}{\tau(n-1)} \max_i L_i + \frac{n(\tau - 1)}{\tau(n-1)} L. \tag{78}$$

Finally,

$$\mathbf{Var}\left[g(y)\right] = \frac{n - \tau}{\tau(n-1)} \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(y)\|^2 - \|\nabla f(y)\|^2 \right). \tag{79}$$

Commentary - I

Let

$$a(\tau) \stackrel{\text{def}}{=} \frac{n - \tau}{\tau(n-1)}, \qquad b(\tau) \stackrel{\text{def}}{=} \frac{n(\tau - 1)}{\tau(n-1)}.$$

Notice that

- $a(\tau) + b(\tau) = 1$ for all $\tau \in \{0, 1, \ldots, n\}$
- a is decreasing, with $a(1) = 1$, $a(n) = 0$
- b is increasing, with $b(1) = 0$, $b(n) = 1$

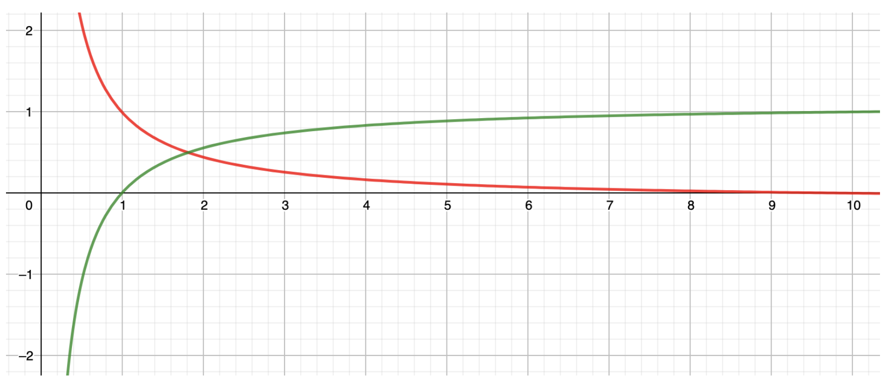

Figure: Functions a and b plotted for $1 \leq \tau \leq n$ and $n = 10$.

Commentary - II

Summary table:

τ	$a(\tau)$	$b(\tau)$	A''	Algorithm
1	1	0	$\max_i L_i$	SGD-US
τ	$\frac{n-\tau}{\tau(n-1)}$	$\frac{n(\tau-1)}{\tau(n-1)}$	$\frac{n-\tau}{\tau(n-1)}\max_i L_i + \frac{n(\tau-1)}{\tau(n-1)} L$	SGD-NICE
n	0	1	L	GD

Key insights:

- For $\tau = 1$, we recover SGD-US, its maximum stepsize, and hence also its rate
- For $\tau = n$, we recover GD, its maximum stepsize, and hence also its rate
- SGD-NICE is therefore a minibatch SGD method that interpolates between SGD-US and GD as τ moves from 1 to n.

Proof of Lemma 45 - I

Unbiasedness. Let χ_i be the random variable defined by

$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}.$$

It is easy to show that

$$\mathbf{E}[\chi_i] = \mathrm{Prob}(i \in S) = \frac{\tau}{n}. \tag{80}$$

Unbiasedness of $g(x)$ now follows via direct computation:

$$\begin{aligned} \mathbf{E}[g(x)] &\overset{(77)}{=} \mathbf{E}\left[\frac{1}{\tau}\sum_{i \in S} \nabla f_i(x)\right] = \mathbf{E}\left[\frac{1}{\tau}\sum_{i=1}^{n} \chi_i \nabla f_i(x)\right] \\ &= \frac{1}{\tau}\sum_{i=1}^{n} \mathbf{E}[\chi_i] \nabla f_i(x) \\ &\overset{(80)}{=} \frac{1}{\tau}\sum_{i=1}^{n} \mathrm{Prob}(i \in S) \nabla f_i(x) \\ &\overset{(80)}{=} \frac{1}{\tau}\sum_{i=1}^{n} \frac{\tau}{n} \nabla f_i(x) \\ &= \nabla f(x). \end{aligned}$$

Proof of Lemma 45 - II

Expected smoothness (i.e., computing constant A''). Fix $x, y \in \mathbb{R}^d$ and let

$$a_i \overset{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \tag{81}$$

Let χ_{ij} be the random variable defined by

$$\chi_{ij} = \begin{cases} 1 & i \in S \text{ and } j \in S \\ 0 & \text{otherwise} \end{cases}.$$

Note that

$$\chi_{ij} = \chi_i \chi_j. \tag{82}$$

Further, it is easy to show that

$$\mathbf{E}\left[\chi_{ij}\right] = \mathrm{Prob}(i \in S, j \in S) = \frac{\tau(\tau-1)}{n(n-1)}. \tag{83}$$

It is easy to check that for any vectors $b_1, \ldots, b_n \in \mathbb{R}^d$ we have the identity

$$\left\|\sum_{i=1}^n b_i\right\|^2 - \sum_{i=1}^n \|b_i\|^2 = \sum_{i \neq j} \langle b_i, b_j \rangle. \tag{84}$$

Proof of Lemma 45 - III

We will use this identity twice in what follows:

$$\begin{aligned}
\mathbf{E}\left[\|g(x) - g(y)\|^2\right] &\overset{(77)}{=} \mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S} \nabla f_i(x) - \frac{1}{\tau}\sum_{i \in S} \nabla f_i(y)\right\|^2\right] \\
&\overset{(81)}{=} \frac{1}{\tau^2}\mathbf{E}\left[\left\|\sum_{i \in S} a_i\right\|^2\right] \\
&= \frac{1}{\tau^2}\mathbf{E}\left[\left\|\sum_{i=1}^n \chi_i a_i\right\|^2\right] \\
&\overset{(84)}{=} \frac{1}{\tau^2}\mathbf{E}\left[\sum_{i=1}^n \|\chi_i a_i\|^2 + \sum_{i \neq j}\langle \chi_i a_i, \chi_j a_j\rangle\right] \\
&\overset{(82)}{=} \frac{1}{\tau^2}\mathbf{E}\left[\sum_{i=1}^n \|\chi_i a_i\|^2 + \sum_{i \neq j}\chi_{ij}\langle a_i, a_j\rangle\right] \\
&= \frac{1}{\tau^2}\sum_{i=1}^n \mathbf{E}\left[\chi_i\right]\|a_i\|^2 + \sum_{i \neq j} \mathbf{E}\left[\chi_{ij}\right]\langle a_i, a_j\rangle.
\end{aligned} \tag{85}$$

Proof of Lemma 45 - IV

Using the formulas (80) and (83) and the decomposition identity (84) again, we can continue:

$$
\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &\overset{(85)}{=} \frac{1}{\tau^2}\left(\frac{\tau}{n}\sum_{i=1}^{n}\|a_i\|^2 + \frac{\tau(\tau-1)}{n(n-1)}\sum_{i\neq j}\langle a_i, a_j\rangle\right) \\
&= \frac{1}{\tau n}\sum_{i=1}^{n}\|a_i\|^2 + \frac{\tau-1}{\tau n(n-1)}\sum_{i\neq j}\langle a_i, a_j\rangle \\
&\overset{(84)}{=} \frac{1}{\tau n}\sum_{i=1}^{n}\|a_i\|^2 + \frac{\tau-1}{\tau n(n-1)}\left(\left\|\sum_{i=1}^{n} a_i\right\|^2 - \sum_{i=1}^{n}\|a_i\|^2\right) \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^{n}\|a_i\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^{n} a_i\right\|^2. \qquad (86)
\end{aligned}
$$

Since f_i is convex and L_i-smooth, we know that

$$\|a_i\|^2 \overset{(81)}{=} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \le 2L_i D_{f_i}(x,y).$$

Since f is convex and L-smooth, we know that

$$\left\|\frac{1}{n}\sum_{i=1}^{n} a_i\right\|^2 \overset{(81)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \le 2L D_f(x,y).$$

Proof of Lemma 45 - V

It only remains to plug these bounds to (86), apply the bound $L_i \le \max_i L_i$ and use the identity $D_f(x,y) = \frac{1}{n}\sum_{i=1}^{n} D_{f_i}(x,y)$:

$$
\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &\overset{(86)}{\le} \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^{n} 2L_i D_{f_i}(x,y) + \frac{n(\tau-1)}{\tau(n-1)}2L D_f(x,y) \\
&\le 2\frac{n-\tau}{\tau(n-1)}\max_i L_i \frac{1}{n}\sum_{i=1}^{n} D_{f_i}(x,y) + 2\frac{n(\tau-1)}{\tau(n-1)}L D_f(x,y) \\
&= 2\frac{n-\tau}{\tau(n-1)}\max_i L_i D_f(x,y) + 2\frac{n(\tau-1)}{\tau(n-1)}L D_f(x,y) \\
&= 2\left(\frac{n-\tau}{\tau(n-1)}\max_i L_i + \frac{n(\tau-1)}{\tau(n-1)}L\right) D_f(x,y).
\end{aligned}
$$

Variance. Using the variance decomposition and then rewriting the second moment using the exact same steps that led to (86), but with $a_i' = \nabla f_i(y)$ instead of a_i, we

Proof of Lemma 45 - VI

arrive at the identity

$$
\begin{aligned}
\mathbf{Var}\left[g(y)\right] &= \mathbf{E}\left[\|g(y)\|^2\right] - \|\mathbf{E}\left[g(y)\right]\|^2 \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^n \|a_i'\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^n a_i'\right\|^2 - \left\|\frac{1}{n}\sum_{i=1}^n a_i'\right\|^2 \\
&= \frac{n-\tau}{\tau(n-1)}\left(\frac{1}{n}\sum_{i=1}^n \|a_i'\|^2 - \left\|\frac{1}{n}\sum_{i=1}^n a_i'\right\|^2\right) \\
&= \frac{n-\tau}{\tau(n-1)}\left(\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(y)\|^2 - \|\nabla f(y)\|^2\right).
\end{aligned}
$$

Convergence of SGD-NICE - I

By combining

- Theorem 33 (main convergence theorem for SGD under the AC assumption),
- Theorem 36 (result that reduced checking the AC assumption to checking expected smoothness: $A = 2A''$ and $C = 2\mathbf{Var}\left[g(x^\star)\right]$) and
- Lemma 45 (computation of expected smoothness constant A''),

we arrive at the complexity result:

Consider the SGD-NICE method with minibatch size $\tau \in \{1, 2, \ldots, n\}$. Assume f_i is convex and L_i-smooth for all i, and that f is μ-convex and L-smooth. Choose any relative error tolerance $0 < \delta < 1$, stepsize $\gamma = \min\left\{\frac{1}{2A''}, \frac{\mu\delta\left\|x^0 - x^\star\right\|^2}{2\mathbf{Var}[g(x^\star)]}\right\}$, where

$$
A'' = \frac{n-\tau}{\tau(n-1)}\max_i L_i + \frac{n(\tau-1)}{\tau(n-1)} L
$$

Convergence of SGD-NICE - II

and

$$\mathbf{Var}\left[g(x^\star)\right] = \frac{n-\tau}{\tau(n-1)} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^\star)\|^2 - \|\nabla f(x^\star)\|^2\right)}_{=\sigma_\star^2}. \tag{87}$$

Then

$$k \geq \max\left\{\frac{2A''}{\mu}, \frac{2\mathbf{Var}\left[g(x^\star)\right]}{\delta\mu^2\left\|x^0 - x^\star\right\|^2}\right\}\log\left(\frac{1}{\delta}\right) \Rightarrow \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq 2\delta\left\|x^0 - x^\star\right\|^2.$$

Moving from relative error δ to absolute error $\varepsilon = 2\delta\left\|x^0 - x^\star\right\|^2$, the above translates to

$$k \geq \max\left\{\frac{2A''}{\mu}, \frac{4\mathbf{Var}\left[g(x^\star)\right]}{\varepsilon\mu^2}\right\}\log\left(\frac{2\left\|x^0 - x^\star\right\|^2}{\varepsilon}\right) \Rightarrow \mathbf{E}\left[\left\|x^k - x^\star\right\|^2\right] \leq \varepsilon.$$

Optimal Minibatch Size - I

Given the above convergence result, we may wish to ask which minibatch size is optimal with respect to the **total complexity** of SGD-NICE, defined as the **product of the number of iterations and the cost of one iteration.** Since

- the number of iterations is $\max\left\{\frac{2A''}{\mu}, \frac{4\mathbf{Var}[g(x^\star)]}{\varepsilon\mu^2}\right\}$ (we ignore the logarithmic factor which does not depend on τ), and
- cost of each iteration is τ,

we arrive at the following **total complexity minimization** problem:

$$\min_{1\leq\tau\leq n} \mathcal{C}(\tau),$$

where

$$\mathcal{C}(\tau) \overset{\text{def}}{=} \frac{2}{\mu(n-1)}\max\left\{\underbrace{(n-\tau)\max_i L_i + n(\tau-1)L}_{\text{increasing linear}}, \underbrace{(n-\tau)\frac{2\sigma_\star^2}{\varepsilon\mu}}_{\text{decreasing linear}}\right\}.$$

Optimal Minibatch Size - II

Observations:

- If $\sigma_\star^2 = 0$ (e.g., in the interpolation regime), then the "decreasing" part is equal to zero, and hence $\mathcal{C}(\tau)$ is an increasing function. So, the optimal minibatch size is

$$\tau^\star = 1.$$

- Further, if $\sigma_\star^2$ is not too large, then the increasing linear function dominates the decerasing linear function on the interval $[1, n]$, and hence the optimal minibatch size is again

$$\tau^\star = 1.$$

This happens for

$$\sigma_\star^2 \leq \frac{\varepsilon\mu \max_i L_i}{2}.$$

Optimal Minibatch Size - III

- Otherwise, the increasing linear and the decreasing linear lines intersect on $(1, n)$, and the optimal minibatch size can be found by computing the intersection:

$$\tau^\star = \frac{n(\theta + L - \max_i L_i)}{\theta + nL - \max_i L_i},$$

where $\theta = \frac{2\sigma_\star^2}{\varepsilon\mu}$.

  - Notice that

$$\sigma_\star^2 = \frac{\varepsilon\mu \max_i L_i}{2} \quad \Rightarrow \quad \theta = \max_i L_i \quad \Rightarrow \quad \tau^\star = \frac{n(\max_i L_i + L - \max_i L_i)}{\max_i L_i + nL - \max_i L_i} = 1.$$

  - Notice that

$$\sigma_\star^2 \to \infty \quad \Rightarrow \quad \tau^\star \to n.$$

Key takeaway: If we care about the total complexity of SGD-NICE, the larger the quantity $\sigma_\star^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|^2 - \|\nabla f(x^\star)\|^2$ is, the larger minibatch size should be chosen. As this quantity grows, the optimal minibatch size approaches n. On the other hand, in the "small" $\sigma_\star^2$ regime (and in particular if $\sigma_\star^2 = 0$), the optimal minibatch size is $\tau^\star = 1$.

Sampling without Replacement
(Independent Sampling)

Sampling without Replacement: Independent Sampling

- Let $p_1, p_2, \ldots, p_n$ be probabilities ($0 < p_i \leq 1$ for all i).
- We do not require these probabilities to add up to 1! So, $\sum_i p_i$ can be anywhere in the interval $(0, n]$.

For each i define a random set as follows:

$$S_i \stackrel{\text{def}}{=} \begin{cases} \{i\} & \text{with probability } p_i \\ \emptyset & \text{with probability } 1 - p_i \end{cases}.$$

We now define a random subset $S \subseteq \{1, 2, \ldots, n\}$ by taking the union of these simple sets

$$S \stackrel{\text{def}}{=} \bigcup_{i=1}^{n} S_i. \tag{88}$$

Define the gradient estimator via

$$g(x) \stackrel{\text{def}}{=} \sum_{i \in S} \frac{1}{np_i} \nabla f_i(x). \tag{89}$$

SGD-IND: The Algorithm

Gradient estimator (89) leads to the following new variant of SGD:

Algorithm 8 SGD-IND

Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, probabilities $0 < p_i \leq 1$ for $i = 1, 2, \ldots, n$
for $k = 0, 1, 2, \ldots$ **do**
  Sample set $S^k = \cup_{i=1}^n S_i^k$, where $S_i^k = \{i\}$ with probability p_i
  $g^k = \sum_{i \in S^k} \frac{1}{np_i} \nabla f_i(x^k)$ obtain a stochastic gradient
  $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

Minibatch Size

Note that S has a random size/cardinality. In such cases, we will use the word minibatch size to refer to the expected cardinality:

$$\tau = \mathbf{E}\left[|S|\right].$$

Note that

$$\mathbf{E}\left[|S|\right] = \mathbf{E}\left[\sum_{i=1}^n |S_i|\right] = \sum_{i=1}^n \mathbf{E}\left[|S_i|\right] = \sum_{i=1}^n 1p_i + 0(1-p_i) = \sum_{i=1}^n p_i. \quad (90)$$

- If we choose $\tau = n$, we must necessarily have $p_i = 1$ for all i, $S \equiv \{1, 2, \ldots, n\}$ and hence we recover the gradient estimator used by GD: $g(x) = \nabla f(x)$.
- If we choose $\tau = 1$, we **do not** recover the gradient estimator used by SGD-NS:

Algorithm	Gradient estimator $g(x)$	Minibatch size τ
SGD-NS	$\frac{1}{np_i}\nabla f_i(x)$	1 (deterministically)
SGD-IND	$\sum_{i \in S} \frac{1}{np_i}\nabla f_i(x)$	1 (in expectation)

SGD-IND: Unbiasedness and Expected Smoothness

Lemma 46

The gradient estimator g defined in (89) is unbiased. If we further assume that f_i is convex and L_i-smooth for all i, and f is L-smooth, then

$$\mathbf{E}\left[\|g(x)-g(y)\|^2\right] \le 2A'' D_f(x,y),$$

where

$$A'' = \frac{\max_i\left(\frac{1}{p_i}-1\right)L_i}{n} + L. \tag{91}$$

Commentary - I

Minibatch size = 1:

Algorithm / probabilities	uniform ($p_i = 1/n$)	nonuniform
SGD-NS	$\max_i L_i$	$\max_i \frac{L_i}{np_i}$
SGD-IND	$L + \frac{n-1}{n}\max_i L_i$	$\frac{\max_i\left(\frac{1}{p_i}-1\right)L_i}{n} + L$

Table: The value of A'' for two variants of SGD under uniform and nonuniform probabilities.

Commentary - II

Note that in the $\tau = 1$ case with uniform probabilities, we have

$$\begin{aligned} A''_{\text{SGD-IND}} &= L + \frac{n-1}{n} \max_i L_i \\ &= \frac{1}{n} nL + \left(1 - \frac{1}{n}\right) \max_i L_i \\ &\geq \frac{1}{n} \max_i L_i + \left(1 - \frac{1}{n}\right) \max_i L_i \\ &= \max_i L_i \\ &= A''_{\text{SGD-US}} \end{aligned}$$

where the inequality follows from Lemma 44(ii). So, SGD-IND has a worse A'' constant than SGD-US, which means its rate is worse.

Minibatch size = n: In the $\tau = n$ case, we recover GD, its maximum stepsize, and hence also its rate since $A'' = L$

Commentary - III

More insights:

- The estimator (89) thus leads to a minibatch SGD method that interpolates between something "similar" to SGD-NS and GD as τ moves from 1 to n.
- Unlike in the case of τ-nice sampling, here we can make use of nonuniform probabilities, which means we can think about constructing **importance sampling for minibatches.**

Proof of Lemma 46 - I

Unbiasedness. As before, let χ_i be the random variable defined by

$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}.$$

It is easy to show that

$$\mathbf{E}[\chi_i] = \mathrm{Prob}(i \in S) = \mathrm{Prob}(i \in S_i) = p_i. \tag{92}$$

Unbiasedness of $g(x)$ now follows via direct computation:

$$\begin{aligned} \mathbf{E}[g(x)] &\overset{(89)}{=} \mathbf{E}\left[\sum_{i\in S} \frac{1}{np_i}\nabla f_i(x)\right] \\ &= \mathbf{E}\left[\sum_{i=1}^n \chi_i \frac{1}{np_i}\nabla f_i(x)\right] \\ &= \sum_{i=1}^n \mathbf{E}[\chi_i]\frac{1}{np_i}\nabla f_i(x) \\ &\overset{(92)}{=} \sum_{i=1}^n \frac{1}{n}\nabla f_i(x) \\ &= \nabla f(x). \end{aligned}$$

Proof of Lemma 46 - II

Expected smoothness (i.e., computing A''). Fix $x, y \in \mathbb{R}^d$ and let

$$a_i \overset{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \tag{93}$$

We will use this identity twice in what follows:

$$\begin{aligned} \mathbf{E}\left[\|g(x)-g(y)\|^2\right] &\overset{(89)}{=} \mathbf{E}\left[\left\|\sum_{i\in S}\frac{1}{np_i}\nabla f_i(x) - \sum_{i\in S}\frac{1}{np_i}\nabla f_i(y)\right\|^2\right] \\ &\overset{(93)}{=} \mathbf{E}\left[\left\|\sum_{i\in S}\frac{a_i}{np_i}\right\|^2\right] \\ &= \mathbf{E}\left[\left\|\sum_{i=1}^n \chi_i\frac{a_i}{np_i}\right\|^2\right] \\ &\overset{(84)}{=} \mathbf{E}\left[\sum_{i=1}^n\left\|\chi_i\frac{a_i}{np_i}\right\|^2 + \sum_{i\neq j}\left\langle \chi_i\frac{a_i}{np_i}, \chi_j\frac{a_j}{np_j}\right\rangle\right] \\ &= \mathbf{E}\left[\sum_{i=1}^n \chi_i\left\|\frac{a_i}{np_i}\right\|^2 + \sum_{i\neq j}\chi_i\chi_j\left\langle \frac{a_i}{np_i}, \frac{a_j}{np_j}\right\rangle\right] \\ &= \sum_{i=1}^n \mathbf{E}[\chi_i]\left\|\frac{a_i}{np_i}\right\|^2 + \sum_{i\neq j}\mathbf{E}[\chi_i\chi_j]\left\langle \frac{a_i}{np_i}, \frac{a_j}{np_j}\right\rangle. \end{aligned} \tag{94}$$

Proof of Lemma 46 - III

Since $\mathbf{E}[\chi_i] = p_i$, and since by independence we have $\mathbf{E}[\chi_i \chi_j] = \mathbf{E}[\chi_i]\,\mathbf{E}[\chi_j] = p_i p_j$, we can further write

$$
\begin{aligned}
\mathbf{E}\left[\|g(x) - g(y)\|^2\right] &= \frac{1}{n^2}\sum_{i=1}^n \frac{1}{p_i}\|a_i\|^2 + \sum_{i \neq j}\left\langle \frac{a_i}{n}, \frac{a_j}{n}\right\rangle \\
&\overset{(84)}{=} \frac{1}{n^2}\sum_{i=1}^n \frac{1}{p_i}\|a_i\|^2 + \left(\left\|\sum_{i=1}^n \frac{a_i}{n}\right\|^2 - \sum_{i=1}^n \left\|\frac{a_i}{n}\right\|^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \left(\frac{1}{p_i} - 1\right)\|a_i\|^2 + \left\|\sum_{i=1}^n \frac{a_i}{n}\right\|^2. \qquad (95)
\end{aligned}
$$

Since f_i is convex and L_i-smooth, we know that

$$\|a_i\|^2 \overset{(93)}{=} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L_i D_{f_i}(x,y).$$

Since f is convex and L-smooth, we know that

$$\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2 \overset{(93)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2L D_f(x,y).$$

Plugging these estimates into (95), and using the identity

$$D_f(x,y) = \frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y),$$

Proof of Lemma 46 - IV

we finally get

$$
\begin{aligned}
\mathbf{E}\left[\|g(x) - g(y)\|^2\right] &\leq \frac{1}{n^2}\sum_{i=1}^n \left(\frac{1}{p_i} - 1\right) 2L_i D_{f_i}(x,y) + 2L D_f(x,y) \\
&\leq 2\frac{\max_i\left(\frac{1}{p_i} - 1\right)L_i}{n}\frac{1}{n}\sum_{i=1}^n D_{f_i}(x,y) + 2L D_f(x,y) \\
&= 2\frac{\max_i\left(\frac{1}{p_i} - 1\right)L_i}{n} D_f(x,y) + 2L D_f(x,y) \\
&= 2\left(\frac{\max_i\left(\frac{1}{p_i} - 1\right)L_i}{n} + L\right) D_f(x,y).
\end{aligned}
$$

Sampling with Replacement

Sampling with Replacement: Multisampling

Let $q_1, q_2, \ldots, q_n$ be probabilities summing up to 1 and let s be the random variable equal to i with probability q_i. Fix a minibatch size $\tau \in \{1, 2, \ldots\}$ and let $s_1, s_2, \ldots, s_\tau$ be independent copies of s. Define the gradient estimator via

$$g(x) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{nq_{s_t}} \nabla f_{s_t}(x). \tag{96}$$

Gradient estimator (96) leads to the following new variant of SGD:

Algorithm 9 SGD-MULTI

Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, positive probabilities $q_1, \ldots, q_n$ summing up to 1, minibatch size $\tau \in \{1, 2, \ldots\}$

for $k = 0, 1, 2, \ldots$ **do**

  Sample τ i.i.d. random variables $s_1^k, \ldots, s_\tau^k$, where each is equal to i with probability q_i

  $g^k = \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{nq_{s_t^k}} \nabla f_{s_t^k}(x^k)$ obtain a stochastic gradient

  $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$

SGD-MULTI: Unbiasedness and Expected Smoothness

Lemma 47

The gradient estimator g defined in (96) *is unbiased. If we further assume that f_i is convex and L_i-smooth for all i, and f is L-smooth, then*

$$\mathbf{E}\left[\|g(x) - g(y)\|^2\right] \le 2A'' D_f(x, y),$$

where

$$A'' = \frac{1}{\tau}\left(\max_i \frac{L_i}{nq_i}\right) + \left(1 - \frac{1}{\tau}\right) L. \qquad (97)$$

Commentary - I

Let

$$a(\tau) \stackrel{\text{def}}{=} \frac{1}{\tau}, \qquad b(\tau) \stackrel{\text{def}}{=} 1 - \frac{1}{\tau}.$$

Notice that

- $a(\tau) + b(\tau) = 1$ for all $\tau \in \{0, 1, \dots\}$
- a is decreasing, with $a(1) = 1$, $a(+\infty) = 0$
- b is increasing, with $b(1) = 0$, $b(+\infty) = 1$

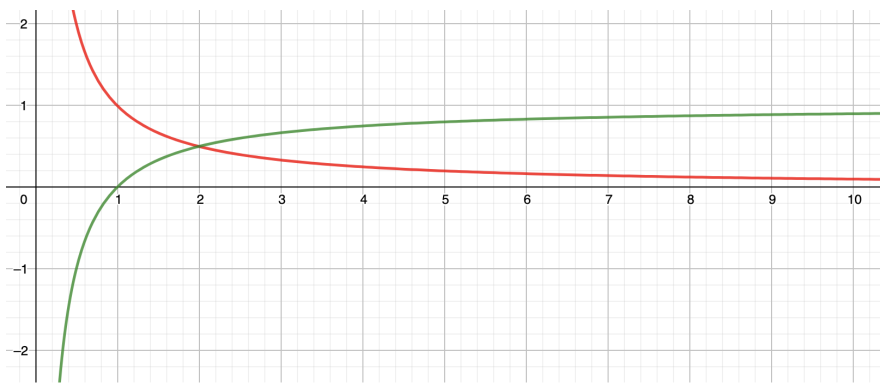

Figure: Functions a and b plotted for $1 \le \tau \le +\infty$.

Commentary - II

Summary table:

τ	$a(\tau)$	$b(\tau)$	A''	Algorithm
1	1	0	$\max_i \frac{L_i}{nq_i}$	SGD-NS
τ	$\frac{1}{\tau}$	$1-\frac{1}{\tau}$	$\frac{1}{\tau}\left(\max_i \frac{L_i}{nq_i}\right)+\left(1-\frac{1}{\tau}\right)L$	SGD-MULTI
$+\infty$	0	1	L	GD

Key insights:

- For $\tau = 1$, we recover SGD-NS, its maximum stepsize, and hence also its rate
- For $\tau = +\infty$, we recover GD, its maximum stepsize, and hence also its rate
- The estimator (77) thus leads to a minibatch SGD method that interpolates between SGD-NS and GD as τ moves from 1 to $+\infty$.

Proof of Lemma 47 - I

Unbiasedness. Unbiasedness of $g(x)$ follows via direct computation:

$$\begin{aligned}
\mathbf{E}[g(x)] &= \mathbf{E}\left[\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{nq_{s_t}}\nabla f_{s_t}(x)\right] \\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}\mathbf{E}\left[\frac{1}{nq_{s_t}}\nabla f_{s_t}(x)\right] \\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}\sum_{i=1}^{n} q_i \frac{1}{nq_i}\nabla f_i(x) \\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) \\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}\nabla f(x) \\
&= \nabla f(x).
\end{aligned}$$

Expected smoothness (i.e., computing A''). Fix $x, y \in \mathbb{R}^d$ and let

$$a_i \overset{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \tag{98}$$

Proof of Lemma 47 - II

Then

$$\begin{aligned}
\mathbf{E}\left[\|g(x)-g(y)\|^2\right] &= \mathbf{E}\left[\left\|\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{nq_{s_t}}\nabla f_{s_t}(x)-\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{nq_{s_t}}\nabla f_{s_t}(y)\right\|^2\right] \\
&\overset{(98)}{=} \frac{1}{\tau^2}\mathbf{E}\left[\left\|\sum_{t=1}^{\tau}\frac{a_{s_t}}{nq_{s_t}}\right\|^2\right] \\
&= \frac{1}{\tau^2}\mathbf{E}\left[\sum_{t=1}^{\tau}\left\|\frac{a_{s_t}}{nq_{s_t}}\right\|^2+\sum_{t\neq u}^{\tau}\left\langle\frac{a_{s_t}}{nq_{s_t}},\frac{a_{s_u}}{nq_{s_u}}\right\rangle\right] \\
&= \frac{1}{\tau^2}\sum_{t=1}^{\tau}\mathbf{E}\left[\left\|\frac{a_{s_t}}{nq_{s_t}}\right\|^2\right]+\frac{1}{\tau^2}\sum_{t\neq u}^{\tau}\mathbf{E}\left[\left\langle\frac{a_{s_t}}{nq_{s_t}},\frac{a_{s_u}}{nq_{s_u}}\right\rangle\right] \qquad (99)
\end{aligned}$$

Proof of Lemma 47 - III

We now separately bound the two terms in (99) by a multiple of the Bregman divergence $D_f(x,y)$. First, we estimate

$$\begin{aligned}
\mathbf{E}\left[\left\|\frac{a_{s_t}}{nq_{s_t}}\right\|^2\right] &= \sum_{i=1}^{n} q_i\left\|\frac{a_i}{nq_i}\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{nq_i}\|a_i\|^2 \\
&\overset{(98)}{=} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{nq_i}\|\nabla f_i(x)-\nabla f_i(y)\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\frac{2L_i}{nq_i}D_{f_i}(x,y) \\
&\leq 2\left(\max_i\frac{L_i}{nq_i}\right)\frac{1}{n}\sum_{i=1}^{n}D_{f_i}(x,y) \\
&= 2\left(\max_i\frac{L_i}{nq_i}\right)D_f(x,y). \qquad (100)
\end{aligned}$$

The first inequality follows from Proposition 17 (iii) since f_i is convex and L_i-smooth, the second inequality follows by bounding $\frac{L_i}{nq_i}\leq\max_i\frac{L_i}{nq_i}$, and the last identity was the subject of an exercise.

Proof of Lemma 47 - IV

Next, since s_t and s_u are independent for $t \neq u$, we have

$$\begin{aligned}
\mathbf{E}\left[\left\langle \frac{a_{s_t}}{nq_{s_t}}, \frac{a_{s_u}}{nq_{s_u}} \right\rangle\right] &= \left\langle \mathbf{E}\left[\frac{a_{s_t}}{nq_{s_t}}\right], \mathbf{E}\left[\frac{a_{s_u}}{nq_{s_u}}\right] \right\rangle \\
&= \left\langle \sum_{i=1}^{n} q_i \frac{a_i}{nq_i}, \sum_{i=1}^{n} q_i \frac{a_i}{nq_i} \right\rangle \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} a_i \right\|^2 \\
&\overset{(98)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq 2LD_f(x, y). \qquad (101)
\end{aligned}$$

Finally, plugging (100) and (101) into (99), we obtain

$$\begin{aligned}
\mathbf{E}\left[\|g(x) - g(y)\|^2\right] &\leq \frac{1}{\tau^2} \sum_{t=1}^{\tau} 2 \left(\max_i \frac{L_i}{nq_i}\right) D_f(x, y) + \frac{1}{\tau^2} \sum_{t \neq u} 2LD_f(x, y) \\
&= 2\left(\frac{1}{\tau}\left(\max_i \frac{L_i}{nq_i}\right) + \left(1 - \frac{1}{\tau}\right) L\right) D_f(x, y),
\end{aligned}$$

as desired.

Exercises

Exercises I

Exercise 40

Show that if S is a τ-nice sampling (i.e., a random subset of $\{1, 2, \ldots, n\}$ of cardinality τ chosen uniformly at random), then

$$\mathrm{Prob}(i \in S) = \frac{\tau}{n}.$$

Exercise 41

Show that if S is a τ-nice sampling (i.e., a random subset of $\{1, 2, \ldots, n\}$ of cardinality τ chosen uniformly at random), then

$$\mathrm{Prob}(i \in S, j \in S) = \frac{\tau(\tau-1)}{n(n-1)}.$$

Exercises II

Exercise 42

Prove identity (84). That is, prove that for any vectors $b_1, \ldots, b_n \in \mathbb{R}^d$,

$$\left\| \sum_{i=1}^{n} b_i \right\|^2 - \sum_{i=1}^{n} \|b_i\|^2 = \sum_{i \neq j} \langle b_i, b_j \rangle.$$

Exercise 43 (Second moment of `SGD-IND`)

Compute $\mathbf{E}\left[\|g(y)\|^2\right]$ for the gradient estimator $g(x)$ defined in (89).