Stochastic Gradient Descent and Mini-batch Gradient Descent

Ruixin Guo

Department of Computer Science
Kent State University

June 12, 2023

Contents

1. Sum of Functions

2. Stochastic Gradient Descent

3. Mini-batch Gradient Descent

Recall

Convex Function: Let $f : C \to \mathbb{R}$ be a convex function where $C \subseteq \mathbb{R}^d$ is a convex set. Then for any $x, y \in C$ and $\lambda \in [0,1]$,
(1) $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$
(2) $f(x) + \langle \nabla f(x), y - x \rangle \leq f(y)$

Lipschitz Smooth: Let $X \subseteq \mathbb{R}^d$ and $L > 0$. A function $f : X \to \mathbb{R}$ is L-smooth if for any $x, y \in X$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Especially, when X is convex, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

Moreover, if $\inf f > -\infty$,

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - \inf f$$

Recall

Strong Convexity: Let $C \subseteq \mathbb{R}^d$ be a convex set and $\mu > 0$. A function $f : C \to \mathbb{R}$ is μ-strongly convex if for any $x, y \in X$ and $\lambda \in [0, 1]$,

$$\mu \frac{t(1-t)}{2} \|x - y\|^2 + f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

If f is also differentiable, then for any $x, y \in C$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$$

Recall

Notation: Let $x \in \mathbb{R}^d$, x_i denotes the i-th element of x, x^k denotes the k-th iteration of x (Note that this is not the k-th power of x).

Gradient Descent: Let $C \subseteq \mathbb{R}^d$ be a convex set and $f : C \to \mathbb{R}$ be a convex function. Let x^* be the minimizer of $\min_{x \in C} f(x)$. Let $x_0 \in C$ be a start point, the gradient descent algorithm makes iteration by $x^{k+1} = x^k - t\nabla f(x^k)$ for $k = 0, 1, 2....$
When f is L-smooth and let the step size $t < 1/L$ be constant, the sequence $\{f(x^k)\}$ will converge to $f(x^*)$ by

$$f(x^k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2tk}$$

which is sublinear in the worst case.

If f is L-smooth and μ-strongly convex, let the step size $t < 1/L$ be constant, the sequence $\{x^k\}$ will converge to x^* by

$$\|x_{k+1} - x^*\|^2 \leq (1 - t\mu)^{k+1} \|x_0 - x^*\|^2$$

which is linear in the worst case.

Smoothness and Convexity

Lemma 0.1: If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth, then for all $x, y \in \mathbb{R}^d$ we have that
(1)

$$\frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle$$

(2)

$$\frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \leq \langle \nabla f(y) - \nabla f(x), y - x\rangle \quad \text{(Co-coercivity)}$$

Proof:
(1) Let's fix $x, y \in \mathbb{R}^d$. For every $z \in \mathbb{R}^d$, we have

$$f(x) - f(y) = f(x) - f(z) + f(z) - f(y)$$

By convexity,

$$f(x) - f(z) \leq \langle \nabla f(x), x - z\rangle$$

By L-smooth,

$$f(z) - f(y) \leq \langle \nabla f(y), z - y\rangle + \frac{L}{2}\|z - y\|^2$$

Thus,

$$f(x) - f(y) \leq \langle \nabla f(x), x - z\rangle + \langle \nabla f(y), z - y\rangle + \frac{L}{2}\|z - y\|^2$$

Now we want to find the tightest upper bound on the right hand side. Taking the derivative with respect to z,

$$\frac{d}{dz}\langle \nabla f(x), x - z\rangle + \langle \nabla f(y), z - y\rangle + \frac{L}{2}\|z - y\|^2 = -\nabla f(x) + \nabla f(y) + L(z - y) = 0$$

We get the minimizer z,

$$z = y - \frac{1}{L}(\nabla f(y) - \nabla f(x))$$

Plugging this into the upper bound, we get

$$\begin{aligned}
f(x) - f(y) &\leq \langle \nabla f(x), x - z\rangle + \langle \nabla f(y), z - y\rangle + \frac{L}{2}\|z - y\|^2 \\
&= \langle \nabla f(x), x - y\rangle + \frac{1}{L}\langle \nabla f(x), \nabla f(y) - \nabla f(x)\rangle \\
&\quad - \frac{1}{L}\langle \nabla f(y), \nabla f(y) - \nabla f(x)\rangle + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \\
&= \langle \nabla f(x), x - y\rangle - \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \\
&= \langle \nabla f(x), x - y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2
\end{aligned}$$

This proves (1).

(2) By interchanging x and y in (1), we have

$$\frac{1}{2L}\|\nabla f(y)-\nabla f(x)\|^2 \leq f(y)-f(x)-\langle \nabla f(x), y-x\rangle$$

$$\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2 \leq f(x)-f(y)-\langle \nabla f(y), x-y\rangle$$

By adding the two inequalities above on both sides, we get

$$\frac{1}{L}\|\nabla f(y)-\nabla f(x)\|^2 \leq \langle \nabla f(y)-\nabla f(x), y-x\rangle$$

Iteration Complexity

Here we give another measure of convergence speed which is closely related to convergence rate and order: iteration complexity. Iteration complexity measures the minimum number of iterations needed to be within a given error.

Definition (Iteration Complexity): Suppose the sequence $\{x^k\}$ converges to x^*, let ϵ be a small enough error, then

$$k(\epsilon) = \min\{k \mid \|x_k - x^*\| < \epsilon\}$$

We often use asymptotic notations (See Appendix 1) to describe $k(\epsilon)$. For example, if $k(\epsilon) = O(\frac{1}{\epsilon})$, then it will take $O(\frac{1}{\epsilon})$ iterations for x_k to approach x^* within ϵ error.

We have the following statements:
(1) $O(\frac{1}{\epsilon})$ complexity means sublinear convergence.

- Let $c > 0$ be a constant, $k = O(\frac{1}{\epsilon}) \Rightarrow k \leq \frac{c}{\epsilon} \Rightarrow \epsilon \leq \frac{c}{k}$, and $\lim_{k\to\infty} \frac{c/(k+1)}{c/k} = 1$.

(2) $O(\log(\frac{1}{\epsilon}))$ complexity means linear convergence.

- Let $c > 0$ be a constant and $a > 1$ be the base of logarithm, then $k = O(\log(\frac{1}{\epsilon})) \Rightarrow k \leq c \log_a \frac{1}{\epsilon} \Rightarrow \epsilon \leq a^{c-k}$, and $\lim_{k\to\infty} \frac{a^{c-(k+1)}}{a^{c-k}} = \frac{1}{a} \in (0, 1)$.

Contents

1. Sum of Functions

2. Stochastic Gradient Descent

3. Mini-batch Gradient Descent

Problem Definition

The stochastic gradient descent can be considered as a problem of minimizing the sum of functions.

Problem (Sum of Functions): We want to minimize a function $f : \mathbb{R}^d \to \mathbb{R}$ which writes as

$$f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$.

Assumption (Sum of Convex): We consider the Problem (Sum of Functions) where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is assumed to be convex. [1]

Assumption (Sum of $L_{\textbf{max}}$ Smooth): We consider the Problem (Sum of Functions) where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is assumed to be L_i-smooth. We note $L_{\text{max}} := \max_{1,2,\dots,n} L_i$ and $L_{\text{avg}} := \frac{1}{n} \sum_{i=1}^{n} L_i$.

[1] Note that f may not be convex and may have multiple local minimums. When f is non-convex, suppose $x^* \in \text{argmin } f$ be one minimizer of f, let $x^* \in C$ where C is a convex set and f is convex on C, we consider solving $\min_{x \in C} f(x)$ instead.

Examples

Consider the Machine Learning case. Let $\boldsymbol{X} = [\boldsymbol{x_1}\ \boldsymbol{x_2}\ \boldsymbol{x_3}] \in \mathbb{R}^{2\times 3}$ be the matrix of training samples where each $\boldsymbol{x_i} = \begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} \in \mathbb{R}^2$ is one training sample. Let $\boldsymbol{y} = [y_1\ y_2\ y_3] \in \mathbb{R}^{1\times 3}$ be the labels such that the sample $\boldsymbol{x_i}$ corresponds to the label y_i. Let $g(\boldsymbol{x};\boldsymbol{\theta}) : \mathbb{R}^2 \to \mathbb{R}$ be a machine learning model where $\boldsymbol{x} \in \mathbb{R}^2$ are the inputs and $\boldsymbol{\theta}$ are the parameters. (Suppose $\boldsymbol{\theta}$ can be of arbitrary size.)

Let's use the Mean Square Error as the Empirical Risk Function:

$$f(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta}) = \frac{1}{3}\sum_{i=1}^{3} f_i(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta})$$

where each f_i is the Loss Function of the i-th sample, defined as

$$f_i(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta}) = \left(g(\boldsymbol{x_i};\boldsymbol{\theta}) - y_i\right)^2$$

We can consider $f(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta})$ as a sum of functions of $f_i (i = 1, 2, 3)$.

Examples

The goal of optimizing machine learning model is to find $\min_{\boldsymbol{\theta}} f(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\theta})$.

Let t be the step size. If using Gradient Descent, the iteration becomes

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^{k} - t\nabla_{\boldsymbol{\theta}} f(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\theta})$$

If using Stochastic Gradient Descent, the iteration becomes

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^{k} - t\nabla_{\boldsymbol{\theta}} f_p(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\theta})$$

where p is randomly chosen from $\{1, 2, 3\}$ in each iteration.

Since

$$\nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\theta}) = 2\left(g(\boldsymbol{x_i}; \boldsymbol{\theta}) - y_i\right)\left[\frac{\partial}{\partial \boldsymbol{\theta}}\, g(\boldsymbol{x_i}; \boldsymbol{\theta})\right]^T$$

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{3}\sum_{i=1}^{3} \nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\theta})$$

It is obvious that the computation cost of $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\theta})$ is 3 times of $\nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\theta})$. In general, if there are n training samples, the computation cost of the gradient of the entire dataset will be n times of the gradient of one training sample.

Expectation and Variance of High-Dimensional Variables

Definition: Given a random variable X in $\mathbb{R}^d$, we note:

- The expectation of X as $\mathbb{E}[X]$.
- The variance of X as $\mathrm{Var}[X] = \mathbb{E}[\|X - \mathbb{E}[X]\|^2]$.

Given two random variables X_1, X_2 in $\mathbb{R}^d$ such that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mathbb{E}[X]$. We note the covariance of X_1, X_2 as $\mathrm{Cov}(X_1, X_2) = \mathbb{E}[\langle X_1 - \mathbb{E}[X], X_2 - \mathbb{E}[X]\rangle]$.

Here $\mathbb{E}[X]$ is a vector, $\mathrm{Var}[X]$ and $\mathrm{Cov}(X_1, X_2)$ are numbers. Moreover, we have

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[\|X - \mathbb{E}[X]\|^2] = \mathbb{E}[\|X\|^2 - 2\langle X, \mathbb{E}[X]\rangle + \|\mathbb{E}[X]\|^2] \\ &= \mathbb{E}[\|X\|^2] - 2\mathbb{E}[\langle X, \mathbb{E}[X]\rangle] + \mathbb{E}[\|\mathbb{E}[X]\|^2] \\ &= \mathbb{E}[\|X\|^2] - 2\|\mathbb{E}[X]\|^2 + \|\mathbb{E}[X]\|^2 = \mathbb{E}[\|X\|^2] - \|\mathbb{E}[X]\|^2\end{aligned}$$

Note that $\|\mathbb{E}[X]\| \neq \mathbb{E}[\|X\|]$.

It is also easy to verity that

$$\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\mathrm{Cov}(X_1, X_2)$$

Interpolation

We can consider f as an interpolation by functions $f_1, f_2, ..., f_n$.

Definition: Consider the Problem (Sum of Functions). We say that interpolation holds if there exists a common x^* such that $f_i(x^*) = \inf f_i$ for all $i = 1, 2, ..., n$. In this case, we say that interpolation holds at x^*.

Lemma 1.1: If interpolation holds at $x^* \in \mathbb{R}^d$, then $x^* \in$ argmin f.

Proof: Since by definition, $x^* \in$ argmin f_i, then for every $x \in \mathbb{R}^d$,

$$f(x^*) = \frac{1}{n} \sum_{i=1}^{n} f_i(x^*) = \frac{1}{n} \sum_{i=1}^{n} \inf f_i \leq \frac{1}{n} \sum_{i=1}^{n} f_i(x) = f(x)$$

which means $x^* \in$ argmin f_i.

Function Noise

Definition (Function Noise): We define the function noise as

$$\Delta_f^* := \inf f - \frac{1}{n}\sum_{i=1}^{n} \inf f_i$$

Lemma 1.2: Consider the Problem (Sum of Functions), we have
(1) $\Delta_f^* \geq 0$.
(2) Interpolation holds if and only if $\Delta_f^* = 0$.

Proof:
(1) Let $x^* \in$ argmin f, then $\frac{1}{n}\sum_{i=1}^{n} \inf f_i$

$$\Delta_f^* = f(x^*) - \frac{1}{n}\sum_{i=1}^{n} \inf f_i \geq \frac{1}{n}\sum_{i=1}^{n} \inf f_i(x^*) = f(x^*) - f(x^*) = 0$$

Function Noise

(2) Let interpolation hold at $x^* \in \mathbb{R}^d$, then $f_i(x^*) = \inf f_i$ for $i = 1, 2, ..., n$. Thus,

$$\Delta_f^* = f(x^*) - \frac{1}{n}\sum_{i=1}^{n} \inf f_i = \frac{1}{n}\sum_{i=1}^{n} \inf f_i(x^*) = f(x^*) - f(x^*) = 0$$

Conversely, if $\Delta_f^* = 0$, then for some $x^* \in$ argmin f,

$$0 = \Delta_f^* = f(x^*) - \frac{1}{n}\sum_{i=1}^{n} \inf f_i = \frac{1}{n}\sum_{i=1}^{n}(f(x^*) - \inf f_i)$$

Since $f(x^*) - \inf f_i \geq 0$, we must have $f(x^*) = \inf f_i$.

Function Noise

Lemma 1.3: If the Assumptions (Sum of $L_{\max}$-smooth) and (Sum of Convex) hold, then
(1) For all $x, y \in \mathbb{R}^d$,

$$\frac{1}{2L_{\max}}\mathbb{E}[\|\nabla f_i(y) - \nabla f_i(x)\|^2] \leq f(y) - f(x) + \langle \nabla f(x), y - x \rangle$$

(2) For every $x \in \mathbb{R}^d$ and every $x^* \in$ argmin f,

$$\frac{1}{2L_{\max}}\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] \leq f(x) - \inf f$$

Proof:
(1) Note that $L_i \leq L_{\max}$. By Lemma 0.1 (1) we have

$$\frac{1}{2L_{\max}}\|\nabla f_i(y) - \nabla f_i(x)\|^2 \leq f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x \rangle$$

Since $f = \frac{1}{n}\sum_{i=1}^n f_i = \mathbb{E} f_i$, we have $\nabla f = \frac{1}{n}\sum_{i=1}^n \nabla f_i = \mathbb{E}\nabla f_i$. Taking the expectation on both sides of the above inequality, we proved (1).
(2) Apply (1) with $x = x^*$ and $y = x$, since $f(x^*) = \inf f$,

$$\frac{1}{2L_{\max}}\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] \leq f(x) - \inf f + \langle \nabla f(x^*), x - x^* \rangle$$

Since $\nabla f(x^*) = 0$, we proved (2).

Gradient Noise

Definition (Gradient Noise): Let Assumption (Sum of L_{max}-smooth) hold. We define the gradient noise as

$$\sigma_f^* := \inf_{x^* \in \text{argmin } f} \text{Var}[\nabla f_i(x^*)]$$

where

$$\text{Var}[\nabla f_i(x^*)] = \mathbb{E}[\|\nabla f_i(x^*) - \mathbb{E}\nabla f_i(x^*)\|^2] = \mathbb{E}[\|\nabla f_i(x^*)\|^2] - \|\nabla f(x^*)\|^2$$

Lemma 1.4: If the Assumption (Sum of L_{max}-smooth) holds, then:
(1) $\sigma_f^* \geq 0$
(2) If the Assumption (Sum of Convex) holds, then $\sigma_f^* = \text{Var}[\nabla f_i(x^*)]$ for every $x^* =$ argmin f.
(3) If interpolation holds then $\sigma_f^* = 0$. This becomes an equivalence if Assumption (Sum of Convex) holds.

Gradient Noise

Proof:

(1) By the definition $\text{Var}[\nabla f_i(x^*)] = \mathbb{E}[\|\nabla f_i(x^*) - \mathbb{E}\nabla f_i(x^*)\|^2] \geq 0.$ since $\sigma_f^* := \inf_{x^* \in \text{argmin } f} \text{Var}[\nabla f_i(x^*)]$, we have $\sigma_f^* \geq 0.$

(2) Let $x^*, x' \in \text{argmin } f$, we will show that $\text{Var}[\nabla f_i(x^*)] = \text{Var}[\nabla f_i(x')]$. By Lemma 1.3 (2) we have

$$\frac{1}{2L_{\text{max}}}\mathbb{E}[\|\nabla f_i(x') - \nabla f_i(x^*)\|^2] \leq f(x) - \inf f = \inf f - \inf f = 0$$

This means $\mathbb{E}[\|\nabla f_i(x') - \nabla f_i(x^*)\|^2] = 0.$ Thus for every $i = 1, 2, ..., n$, $\|\nabla f_i(x') - \nabla f_i(x^*)\| = 0 \Rightarrow \nabla f_i(x') = \nabla f_i(x^*).$

Therefore $\text{Var}[\nabla f_i(x^*)] = \text{Var}[\nabla f_i(x')]$.

(3) If interpolation holds, there exists a $x^* \in \mathbb{R}^d$ such that $x^* \in f_i$. By Fermat's Theorem [2], this implies that $\nabla f_i(x^*) = 0$ and $\nabla f_i(x^*) = 0.$ Thus $\text{Var}[\nabla f_i(x^*)] = \mathbb{E}[\|\nabla f_i(x^*) - \nabla f(x^*)\|^2] = 0 \Rightarrow \sigma_f^* = 0.$

Conversely, if $\sigma_f^* = 0$, then there exists $x^* \in \text{argmin } f$ such that all $\text{Var}[\nabla f_i(x^*)]$ are equivalent. Since $x^* \in \text{argmin } f$, $\nabla f(x^*) = 0.$ Since $\nabla f(x^*) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*)$, we know that $\nabla f_i(x^*) = 0$ for $i = 1, 2, ..., n$. Thus x^* is the common minimizer of all f_i, the interpolation holds.

[2] Let $f : X \to \mathbb{R}$ where X is an open set, suppose $x_0 \in X$ is a point where f has local extremum. If f is differentiable at x_0, then $f'(x_0) = 0.$

Variance Transfer

Lemma 1.5 (Variance Transfer: Function Noise): If Assumption (Sum of L_{max}-smooth) holds, then for all $x \in \mathbb{R}^d$ we have

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2L_{\mathsf{max}}(f(x) - \inf f) + 2L_{\mathsf{max}}\Delta_f^*$$

Proof:
Remember that if f is L-smooth then

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - \inf f$$

Let $x \in \mathbb{R}^d$ and $x^* \in \text{argmin } f$, then

$$\begin{aligned}\|\nabla f_i(x)\|^2 &\leq 2L_{\mathsf{max}}(f_i(x) - \inf f_i) \\ &= 2L_{\mathsf{max}}(f_i(x) - f_i(x^*)) + 2L_{\mathsf{max}}(f_i(x^*) - \inf f_i)\end{aligned}$$

Taking expectation on both sides, we get

$$\begin{aligned}\mathbb{E}[\|\nabla f_i(x)\|^2] &\leq 2L_{\mathsf{max}}(f(x) - \inf f) + 2L_{\mathsf{max}}(\inf f - \frac{1}{n}\sum_{i=1}^{n} \inf f_i) \\ &= 2L_{\mathsf{max}}(f(x) - \inf f) + 2L_{\mathsf{max}}\Delta_f^*\end{aligned}$$

Variance Transfer

Lemma 1.6 (Variance Transfer: Gradient Noise): If Assumption (Sum of L_{max}-smooth) and (Sum of Convex) holds, then for all $x \in \mathbb{R}^d$ we have

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 4L_{\mathsf{max}}(f(x) - \inf f) + 2\sigma_f^*$$

Proof: Let $x^* \in$ argmin f, then[3]

$$\|\nabla f_i(x)\|^2 \leq 2\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + 2\|\nabla f_i(x^*)\|^2$$

Taking expectation on both sides. By Lemma 1.3 (2) we have

$$\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] \leq 2L_{\mathsf{max}}(f(x) - \inf f)$$

By Lemma 1.4 (2), $\sigma_f^* = \mathsf{Var}[\nabla f_i(x^*)]$ for every $x^* \in$ argmin f since Assumption (Sum of Convex), thus

$$\mathbb{E}[\|\nabla f_i(x^*)\|^2] = \mathsf{Var}[\nabla f_i(x^*)] + \|\nabla f(x^*)\|^2 = \mathsf{Var}[\nabla f_i(x^*)] = \sigma_f^*$$

[3] This is because $\|a\|^2 \leq 2\|a-b\|^2 + 2\|b\|^2 \iff \|a-2b\|^2 \geq 0$. Note that the square of Euclidean Norm does not satisfy triangle inequality.

Summary

- Our goal is to solve the Sum of Functions problem: minimizing a function $f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, where each f_i is assumed to be convex and L_i-smooth. Let $L_{\mathsf{max}} = \max_i L_i$, we say f is L_{max}-smooth.
- In machine learning, we can consider each f_i as the loss of one training sample; and f as the empirical risk, i.e., the average loss of all samples.
- Gradient Descent computes ∇f in each iteration. Stochastic Gradient Descent computes ∇f_i in each iteration, where i is randomly chosen from $\{1, 2, ..., n\}$. The computation cost of ∇f is n times of ∇f_i.
- If f and each f_i have a common minimizer x^*, we say the interpolation holds at x^*. Both function noise Δ_f^* and gradient noise σ_f^* will be 0 if and only if the interpolation holds. Otherwise, both Δ_f^* and σ_f^* will be non-zero. The values of Δ_f^* and σ_f^* only depends on f.

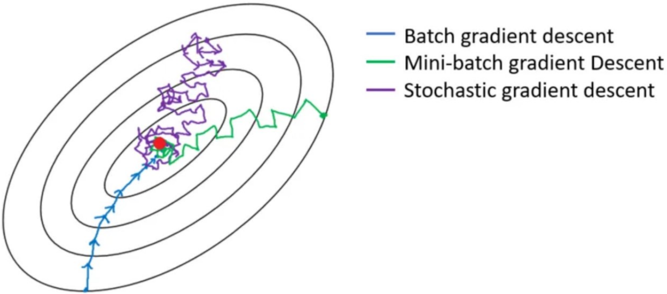
Contents

1. Sum of Functions

2. Stochastic Gradient Descent

3. Mini-batch Gradient Descent

Comparison of GD, SGD and Mini-batch GD



- (Batch) Gradient Decent moves towards the minimum in each iteration.
- Stochastic Gradient Descent may not move towards the minimum in one iteration since the direction is randomly chosen, but it will converges to the minimum in expectation.
- Mini-batch Gradient Descent is a trade-off between the above two. It converges to to the minimum in expectation but with less fluctuation in the movement.

Stochastic Gradient Descent

Algorithm (SGD): Consider Problem (Sum of Functions). Let $x^0 \in \mathbb{R}^d$ be the initial point, and let $\{t_k\}$ be a sequence of step sizes where each $t_k > 0$. The Stochastic Gradient Decent (SGD) algorithm is given by the iterates $\{x^k\}_{k\in\mathbb{N}}$ where:

$$i_k \in \{1, 2, ..., n\} \qquad \text{Sampled with probability } \frac{1}{n}$$

$$x^{k+1} = x^k - t_k \nabla f_{i_k}(x^k)$$

Remark (Unbiased Estimator of Gradient): An important feature of SGD is that each iteration follows a random direction $-\nabla f_{i_k}(x^k)$, which is a unbiased estimator of $-\nabla f(x^k)$, i.e.,

$$\mathbb{E}[\nabla f_i(x^k)|x^k] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^k) = \nabla f(x^k)$$

Here when x^k is given, we consider $\nabla f_i(x^k)$ as a random variable, all possible choices of this random variable are $\nabla f_1(x^k), \nabla f_2(x^k), ..., \nabla f_n(x^k)$, and each with $1/n$ probability.

Convergence Analysis of SGD

Theorem 2.1: Let Assumptions (Sum of L_{max}-smooth) and (Sum of Convex) hold. Consider a sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by SGD, with the step size in each iteration satisfying $0 < t_k < \frac{1}{2L_{\text{max}}}$, then

$$\mathbb{E}[f(\bar{x}^k) - \inf f] \leq \frac{\|x^0 - x^*\|^2}{2\sum_{i=0}^{k-1} t_i(1-2t_iL_{\text{max}})} + \frac{\sum_{i=0}^{k-1} t_i^2}{\sum_{i=0}^{k-1} t_i(1-2t_iL_{\text{max}})}\sigma_f^*$$

where $\bar{x}^k = \sum_{i=0}^{k-1} p_{i,k}x^k$, with $p_{i,k} = \frac{t_i(1-2t_iL_{\text{max}})}{\sum_{j=0}^{k-1} t_j(1-2t_jL_{\text{max}})}$.

Proof: Let $x^* \in \operatorname{argmin} f$, so we have $\sigma_f^* = \text{Var}[\nabla f_i(x^*)]$. We will note $\mathbb{E}_k[\cdot]$ instead of $\mathbb{E}_k[\cdot|x^k]$ for simplicity. First, let's observe the behavior of $\|x^{k+1} - x^*\|^2$

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - t_k\nabla f_i(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2t_k\langle\nabla f_i(x^k), x^k - x^*\rangle + t_k^2\|\nabla f_i(x^k)\|^2\end{aligned}$$

Taking the expectation conditioned on x^k on both sides. Since by convexity,

$$\langle\nabla f(x^k), x^k - x^*\rangle \leq f(x^*) - f(x^k) = \inf f - f(x^k)$$

And by Lemma 1.6,

$$\mathbb{E}_k[\|\nabla f_i(x^k)\|^2] \leq 4L_{\text{max}}(f(x^k) - \inf f) + 2\sigma_f^*$$

Thus,

$$\begin{aligned}\mathbb{E}_k[\|x^{k+1}-x^*\|^2] &= \|x^k-x^*\|^2 + 2t_k\langle\nabla f(x^k), x^*-x^k\rangle + t_k^2\mathbb{E}_k[\|\nabla f_i(x^k)\|^2] \\ &\le \|x^k-x^*\|^2 + 2t_k(\inf f - f(x^k)) + 4t_k^2 L_{\mathsf{max}}(f(x^k)-\inf f) + 2t_k^2\sigma_f^* \\ &= \|x^k-x^*\|^2 + 2t_k(2t_kL_{\mathsf{max}}-1)(f(x^k)-\inf f) + 2t_k^2\sigma_f^*\end{aligned}$$

Rearranging the above inequality,

$$2t_k(2t_kL_{\mathsf{max}}-1)(f(x^k)-\inf f) \le \|x^k-x^*\|^2 - \mathbb{E}_k[\|x^{k+1}-x^*\|^2] + 2t_k^2\sigma_f^*$$

Taking expectation on both sides (Here we consider x^k as a random variable),

$$\begin{aligned}2t_k(2t_kL_{\mathsf{max}}-1)\mathbb{E}[f(x^k)-\inf f] &\le \mathbb{E}[\|x^k-x^*\|^2] - \mathbb{E}[\mathbb{E}_k[\|x^{k+1}-x^*\|^2]] + 2t_k^2\sigma_f^* \\ &= \mathbb{E}[\|x^k-x^*\|^2] - \mathbb{E}\|x^{k+1}-x^*\|^2 + 2t_k^2\sigma_f^*\end{aligned}$$

This is because of Law of Total Expectation, $\mathbb{E}\|x^{k+1}-x^*\|^2 = \mathbb{E}[\mathbb{E}_k[\|x^{k+1}-x^*\|^2]]$. Note that $\mathbb{E}\|x^0-x^*\|^2 = \|x^0-x^*\|^2$ because x^0 is fixed. Summing over $i=0,1,...,k-1$, we get

$$2\sum_{i=0}^{k-1} t_i(2t_iL_{\mathsf{max}}-1)\mathbb{E}[f(x^i)-\inf f] \le \|x^0-x^*\|^2 - \mathbb{E}\|x^k-x^*\|^2 + 2\sigma_f^*\sum_{i=0}^{k-1} t_i^2$$

Since $\mathbb{E}\|x^k - x^*\|^2 \geq 0$, dividing both sides by $2\sum_{i=0}^{k-1} t_i(2t_i L_{\max} - 1)$, we get

$$\mathbb{E}\left[\sum_{i=0}^{k-1} \frac{t_i(2t_i L_{\max} - 1)}{2\sum_{i=0}^{k-1} t_i(2t_i L_{\max} - 1)}(f(x^i) - \inf f)\right] \leq \frac{\|x^0 - x^*\|^2}{2\sum_{i=0}^{k-1} t_i(2t_i L_{\max} - 1)} + \frac{\sigma_f^* \sum_{i=0}^{k-1} t_i^2}{\sum_{i=0}^{k-1} t_i(2t_i L_{\max} - 1)}$$

Define

$$p_{i,k} := \frac{t_i(1 - 2t_i L_{\max})}{\sum_{j=0}^{k-1} t_j(1 - 2t_j L_{\max})},\ \bar{x}^k = \sum_{i=0}^{k-1} p_{i,k} x^i$$

such that $\sum_{i=0}^{k-1} p_{i,k} = 1$. By Jensen's Inequality,

$$\sum_{i=0}^{k-1} p_{i,k} f(x^i) \geq f(\sum_{i=0}^{k-1} p_{i,k} x^i) = f(\bar{x}^k)$$

Therefore[4],

$$\mathbb{E}[f(\bar{x}^k) - \inf f] \leq \frac{\|x^0 - x^*\|^2}{2\sum_{i=0}^{k-1} t_i(1 - 2t_i L_{\max})} + \frac{\sum_{i=0}^{k-1} t_i^2}{\sum_{i=0}^{k-1} t_i(1 - 2t_i L_{\max})}\sigma_f^*$$

[4] Here if we use $\min_{0\leq i\leq k-1} f(x^i)$ to replace $f(\bar{x}^k)$, the inequality still holds since $\sum_{i=0}^{k-1} p_{i,k} f(x^i) \geq \min_{0\leq i\leq k-1} f(x^i)$

Convergence Rate of SGD

Since

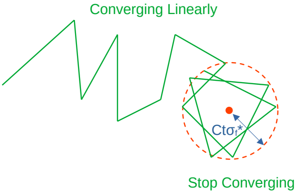
$$\mathbb{E}[f(\bar{x}^k) - \inf f] \leq \frac{\|x^0 - x^*\|^2}{2\sum_{i=0}^{k-1} t_i(1 - 2t_i L_{\mathsf{max}})} + \frac{\sum_{i=0}^{k-1} t_i^2}{\sum_{i=0}^{k-1} t_i(1 - 2t_i L_{\mathsf{max}})}\sigma_f^*$$

We discuss the convergence of SGD in some cases:
Case 2.2.1: The step size is fixed, i.e., we choose $0 < t < \frac{1}{2L_{\mathsf{max}}}$ such that $t_k = t$ for $k = 0, 1, 2, \ldots$. In this case, we can write

$$\mathbb{E}[f(\bar{x}^k) - \inf f] \leq C\frac{\|x^0 - x^*\|^2}{2kt} + Ct\sigma_f^*$$

where $C = 1/(1 - 2tL_{\mathsf{max}})$ is a constant. When $k \in \infty$, $\mathbb{E}[f(\bar{x}^k) - \inf f] \leq Ct\sigma_f^*$.

- If interpolation holds for f, $\sigma_f^* = 0$. $f(\bar{x}^k)$ will converge and the iteration complexity is $O(\frac{1}{\epsilon})$, which is sublinear.
- If interpolation does not hold for f, then $f(\bar{x}^k)$ will not converge. In fact, it will stop converge at a distance $Ct\sigma_f^*$ to $\inf f$. See the right figure.



Convergence Rate of SGD

Case 2.2.2: If $\sigma_f^* > 0$, we can choose step sizes $\{t_k\}_{k \in \mathbb{N}}$ such that $\frac{\sum_{i=0}^{k-1} t_i^2}{\sum_{i=0}^{k-1} t_i} \to 0$ when $k \to \infty$. To do this, let $t_0 = \alpha$ where $\alpha \in (0, \frac{1}{2L_{\max}})$ is a constant. And let $t_i = \frac{\alpha}{i+1}$ for $i = 1, 2, 3....$ Thus,

$$\mathbb{E}[f(\bar{x}^k) - \inf f] \leq C \frac{\|x^0 - x^*\|^2}{2 \sum_{i=1}^{k} \frac{1}{i}} + C\alpha^2 \frac{\sum_{i=1}^{k} \frac{1}{i^2}}{\sum_{i=1}^{k} \frac{1}{i}} \sigma_f^*$$

where $C = \frac{1}{\alpha(1-2\alpha L_{\max})}$.

Consider the term $\frac{\sum_{i=1}^{k} \frac{1}{i^2}}{\sum_{i=1}^{k} \frac{1}{i}}$. It is known that

- $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6} = \Theta(1)$ (See [5])
- $\sum_{i=1}^{k} \frac{1}{i} = \Theta(\log k)$ (See Appendix 2)

Thus, after k iterations, the error is $\Theta(\frac{1}{\log k})$. Let the error be ϵ, then $\frac{1}{\log k} = \epsilon \iff k = e^{1/\epsilon}$ (suppose the base of logarithm is e), which means the iteration complexity is $\Theta(e^{1/\epsilon})$. This is intolerably slow!

[5] https://en.wikipedia.org/wiki/Basel_problem

Convergence Rate of SGD

Case 2.2.3: In Case 2.2.1 and 2.2.2, we set the step sizes $\{t_k\}_{k\in\mathbb{N}}$ to see how ϵ changes with k. Now let's use another strategy: given an ϵ, we choose step sizes $\{t_k\}_{k\in\mathbb{N}}$ to minimize the number of iterations k.

Still consider the fixed step size. By Case 2.2.1 we have

$$\mathbb{E}[f(\bar{x}^k) - \inf f] \leq \frac{\|x^0 - x^*\|^2}{2kt(1-2tL_{\mathsf{max}})} + \frac{t}{(1-2tL_{\mathsf{max}})}\sigma_f^*$$

Let $A = \frac{\|x^0 - x^*\|^2}{2}$, $B = 2L_{\mathsf{max}}$ and $C = \sigma_f^*$, we want

$$\frac{A}{kt(1-Bt)} + \frac{Ct}{1-Bt} \leq \epsilon \quad \Longrightarrow \quad \frac{A}{t\epsilon - (B\epsilon + C)t^2} \leq k$$

When $t = \frac{\epsilon}{2(B\epsilon+C)}$, we have $\min\{\frac{A}{t\epsilon-(B\epsilon+C)t^2}\} = \frac{4AB}{\epsilon} + \frac{4AC}{\epsilon^2}$. Thus $k = \Omega(\frac{1}{\epsilon^2})$.

Note that the above inequality assumes $\frac{\epsilon}{2(B\epsilon+C)} < \frac{1}{2L_{\mathsf{max}}}$. If $\frac{\epsilon}{2(B\epsilon+C)} > \frac{1}{2L_{\mathsf{max}}}$, we let $t = \frac{1}{2L_{\mathsf{max}}}$. Since $t = \frac{\epsilon}{2(B\epsilon+C)}$ is the minimizer of $\frac{A}{t\epsilon-(B\epsilon+C)t^2}$, any other value of t will only make $\frac{A}{t\epsilon-(B\epsilon+C)t^2}$ greater than $\frac{4AB}{\epsilon} + \frac{4AC}{\epsilon^2}$. So $k = \Omega(\frac{1}{\epsilon^2})$ still holds.

Therefore, when $t = \min\{\frac{\epsilon}{2(2L_{\mathsf{max}}\epsilon+\sigma_f^*)}, \frac{1}{2L_{\mathsf{max}}}\}$, the iteration complexity is $\Omega(\frac{1}{\epsilon^2})$.

Strong Convexity Case

Theorem 2.3: Let Assumption (Sum of L_{max}) and (Sum of Convex) hold, and assume further that f is μ-strongly convex. Consider the $\{x^k\}_{k\in\mathbb{N}}$ sequence generated by the SGD algorithm with a constant step size satisfying $0 \leq t \leq \frac{1}{2L_{\text{max}}}$. Then for $t \geq 0$,

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - t\mu)^k \|x^0 - x^*\|^2 + \frac{2t}{\mu}\sigma_f^*$$

Proof: Let $x^* \in$ argmin f, so that $\sigma_f^* = \text{Var}[\nabla f_i(x^*)]$. We will note $\mathbb{E}_k[\cdot]$ instead of $\mathbb{E}_k[\cdot|x^k]$ for simplicity. First, let's observe the behavior of $\|x^{k+1} - x^*\|^2$

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - t\nabla f_i(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2t\langle \nabla f_i(x^k), x^k - x^*\rangle + t^2\|\nabla f_i(x^k)\|^2 \end{aligned}$$

Taking expectation conditioned on x^k on both sides. By strong convexity we have

$$\langle \nabla f(x^k), x^k - x^*\rangle \geq f(x^k) - f(x^*) + \frac{\mu}{2}\|x^k - x^*\|^2$$

And by Lemma 1.6,

$$\mathbb{E}_k[\|\nabla f_i(x^k)\|^2] \leq 4L_{\text{max}}(f(x^k) - \inf f) + 2\sigma_f^*$$

Thus,

$$\begin{aligned}\mathbb{E}_k[\|x^{k+1}-x^*\|^2] &= \|x^k-x^*\|^2 - 2t\langle \nabla f(x^k), x^k-x^*\rangle + t^2\mathbb{E}_k[\|\nabla f_i(x^k)\|^2] \\ &\leq \|x^k-x^*\|^2 - t\mu\|x^k-x^*\|^2 - 2t(f(x^k)-\inf f) \\ &\quad + 4t^2L_{\mathsf{max}}(f(x^k)-\inf f) + 2t^2\sigma_f^* \\ &= (1-t\mu)\|x^k-x^*\|^2 + 2t(2t_kL_{\mathsf{max}}-1)(f(x^k)-\inf f) + 2t^2\sigma_f^* \\ &\leq (1-t\mu)\|x^k-x^*\|^2 + 2t^2\sigma_f^*\end{aligned}$$

where we use the facts that $2t_kL_{\mathsf{max}}-1<0$ and $f(x^k)-\inf f\geq 0$. Taking expectation with respect to x^k on both sides,

$$\mathbb{E}[\|x^{k+1}-x^*\|^2] \leq (1-t\mu)\mathbb{E}[\|x^k-x^*\|^2] + 2t^2\sigma_f^*$$

Recursively applying the above inequality, we get

$$\begin{aligned}\mathbb{E}[\|x^k-x^*\|^2] &\leq (1-t\mu)^k\|x^0-x^*\|^2 + 2\sum_{i=0}^{k-1}(1-t\mu)^i t^2\sigma_f^* \\ &\leq (1-t\mu)^k\|x^0-x^*\|^2 + \frac{2t}{\mu}\sigma_f^*\end{aligned}$$

Convergence Rate of SGD with Strong Convexity

Theorem 2.4: Suppose f is L_{max}-smooth and μ-strongly convex. Let

$$t = \min\left\{\frac{\epsilon}{2}\frac{\mu}{2\sigma_f^*}, \frac{1}{2L_{\mathsf{max}}}\right\}$$

then

$$k \geq \max\left\{\frac{1}{\epsilon}\frac{4\sigma_f^*}{\mu^2}, \frac{2L_{\mathsf{max}}}{\mu}\right\}\log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon}\right) \quad \Longrightarrow \quad \|x^k - x^*\| \leq \epsilon$$

which means the iteration complexity of SGD is $\Omega(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$.

Proof: By Lemma A.2 (See Appendix 3), let $A = \frac{2\sigma_f^*}{\mu}$, $C = 2L_{\mathsf{max}}$ and $\alpha_0 = \|x^0 - x^*\|^2$, we get the result.[6]

[6] Note that here we choose t by assuming that $At \leq \frac{\epsilon}{2}$. Unlike what we showed in Case 2.2.3, this t does not minimize k. In fact, it is unlikely to find an analytical solution of the minimizer t because $(1 - t\mu)^k\alpha_0 + At = \epsilon \Longrightarrow k(t) = \frac{\log(\epsilon - At)/\alpha_0}{\log(1-t\mu)}$, and $k'(t) = 0$ is a transcendental equation.

Comparison of GD and SGD

Remember that when f is L-smooth and μ-strongly convex, GD converges linearly. Using this setting for GD, and using the settings of Theorem 2.4 for SGD. We have

Algorithm	# Iterations	Cost per iteration[7]	Total cost
GD	$O(\log(\frac{1}{\epsilon}))$	$O(nd)$	$O(nd\log(\frac{1}{\epsilon}))$
SGD	$\Omega(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$	$O(d)$	$\Omega(\frac{d}{\epsilon}\log(\frac{1}{\epsilon}))$

This shows when n is large, SGD will have lower total cost than GD.

[7] The computation of $\nabla f_i(x)$ costs $O(d)$ when $x \in \mathbb{R}^d$, and $\nabla f(x)$ costs $O(nd)$.

Contents

1 Sum of Functions

2 Stochastic Gradient Descent

3 Mini-batch Gradient Descent

Mini-batch Function

In practice, the Mini-batch Gradient Descent (Mini-batch GD) is more often used than GD and SGD for large datasets. While GD computes the gradient of all f_is and SGD computes the gradient of one f_i, Mini-batch computes the gradient of a small batch of f_is.

Definition: Let $B \subset \{1, 2, ..., n\}$, and the mini-batch function f_B is defined as

$$f_B(x^k) := \frac{1}{|B|} \sum_{i \in B} f_i(x^k)$$

Taking derivative on both sides. The gradient of f_B is defined as

$$\nabla f_B(x^k) := \frac{1}{|B|} \sum_{i \in B} \nabla f_i(x^k)$$

Mini-batch Gradient Descent

The Mini-batch GD algorithm simply replaces ∇f_i in the SGD algorithm by ∇f_B. Note that ∇f_B is also stochastic.

Algorithm (Mini-batch GD): Let $x^0 \in \mathbb{R}^d$ be the initial point, and let $\{t_k\}$ be a sequence of step sizes where each $t_k > 0$. Let $b \in \{1, 2, ..., n\}$ be the batch size. The Mini-batch Gradient Decent (Mini-batch GD) algorithm is given by the iterates $\{x^k\}_{k \in \mathbb{N}}$ where:

$$\begin{aligned} B_k &\subset \{1, 2, ..., n\} && \text{Sampled uniformly among sets of size } b \\ x^{k+1} &= x^k - t_k \nabla f_{B_k}(x^k) \end{aligned}$$

Properties of Mini-batch Function

Remark: Let $B \subset \{1, 2, ..., n\}$ and $|B| = b < n$, we know that B is formed by choosing b objects in n objects. Thus B has $\binom{n}{b} = \frac{n!}{(n-b)!b!}$ possibilities. The expectation of ∇f_B can be written as

$$\mathbb{E}[\nabla f_B(x^k)] = \frac{1}{\binom{n}{b}} \sum_{\substack{B \subset \{1,2,\ldots,n\} \\ |B|=b}} \nabla f_B(x^k)$$

In fact, we have $\mathbb{E}[\nabla f_B(x^k)] = \nabla f(x^k)$, because

$$\mathbb{E}[\nabla f_B(x^k)] = \frac{1}{\binom{n}{b}} \sum_{\substack{B \subset \{1,2,\ldots,n\} \\ |B|=b}} \left(\frac{1}{|B|} \sum_{B} \mathbb{E}[\nabla f_i(x^k)] \right)$$

where $\mathbb{E}[\nabla f_i(x^k)] = \nabla f(x^k)$.

Properties of Mini-batch Function

Definition 3.1 ($\mathcal{L}_b$-smooth in expectation): Let Assumption (Sum of L_{max}-smooth) hold, and let $b \in \{1, 2, ..., n\}$. We say that f is $\mathcal{L}_b$-smooth in expectation if for all $x, y \in \mathbb{R}^d$,

$$\frac{1}{2\mathcal{L}_b}\|\nabla f_B(y) - \nabla f_B(x)\|^2 \leq f(y) - f(x) + \langle \nabla f(x), y - x\rangle$$

Note that this is similar to Lemma 1.3 (1).

Definition 3.2 (Mini-batch gradient noise): Let Assumption (Sum of L_{max}-smooth) hold, and let $b \in \{1, 2, ..., n\}$. We define the mini-batch gradient noise as

$$\sigma_b^* := \inf_{x^* \in \text{argmin } f} \text{Var}[\nabla f_B(x^*)]$$

Lemma 3.3: Let Assumptions (Sum of L_{max}-smooth) and (Sum of Convex) hold. It follows that

$$\mathbb{E}_k[\|\nabla f_B(x^k)\|^2] \leq 4\mathcal{L}_b(f(x^k) - \inf f) + 2\sigma_b^*$$

Proof: This Lemma is exactly the Lemma 1.6 except for replacing L_{max} and σ_f^* by $\mathcal{L}_b$ and σ_b^*. The proof will be the same except for applying $\mathcal{L}_b$ from Definition 3.1 and σ_b^* from Definition 3.2.

Properties of Mini-batch Function

Lemma 3.4: Let Assumptions (Sum of L_{max}-smooth) and (Sum of Convex) hold, and f is $\mathcal{L}_b$-smooth in expectation. The mini-batch gradient noise can be computed via

$$\sigma_b^* = \frac{n-b}{b(n-1)}\sigma_f^*$$

And $\mathcal{L}_b$ can be bounded by

$$\mathcal{L}_b \leq \frac{n(b-1)}{b(n-1)}L_{\text{avg}} + \frac{n-b}{b(n-1)}L_{\text{max}}$$

Proof:
(1) Since Assumption (Sum of Convex) holds, Var$[\nabla f_i(x^*)]$ is identical for all $x^* \in$ argmin f by Lemma 1.4 (2), we can ignore the $\inf$ notation in σ_b^* and σ_f^*. Thus $\sigma_b^* =$ Var$[\nabla f_B(x^*)]$ and $\sigma_f^* =$ Var$[\nabla f_i(x^*)]$.

Now we can consider $\nabla f_1(x^*), \nabla f_2(x^*), ..., \nabla f_n(x^*)$ as a finite population of size n. Then $\nabla f_i(x^*)$ is a sample of size 1 from the population, and $\nabla f_B(x^*)$ is the mean of a sample of size b. By Lemma A.3 (See Appendix 4), we have

$$\sigma_b^* = \frac{n-b}{b(n-1)}\sigma_f^*$$

Properties of Mini-batch Function

(2) First we prove if each f_i is L_i-smooth, then f is L_{avg}-smooth where $L_{\mathsf{avg}} = \frac{1}{n}\sum_{i=1}^{n} L_i$. Consider the definition of Lipschitz smooth: if f is L-smooth then $\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|$ By triangular inequality we have

$$\|\nabla f(y) - \nabla f(x)\| = \|\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(y) - \nabla f_i(x)\| \le \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(y) - \nabla f_i(x)\|$$
$$\le \frac{1}{n}\sum_{i=1}^{n} L_i\|y - x\| = L_{\mathsf{avg}}\|y - x\|$$

Thus f is L_{avg}-smooth. And by Lemma 0.1 (1),

$$\|\nabla f(y) - \nabla f(x)\|^2 \le 2L_{\mathsf{avg}}(f(y) - f(x) - \langle \nabla f(x), y - x\rangle)$$

Since f is L_{max}-smooth, by Lemma 1.3 (1),

$$\mathbb{E}[\|\nabla f_i(y) - \nabla f_i(x)\|^2] \le 2L_{\mathsf{max}}(f(y) - f(x) + \langle \nabla f(x), y - x\rangle)$$

Properties of Mini-batch Function

Similar to the proof of Lemma A.3, we let $Z_1, Z_2, ..., Z_n$ be random variables where each $Z_i \in \{0, 1\}$. $Z_i = 1$ means $i \in B$ and $Z_i = 0$ means $i \notin B$. Z_i satisfy a distribution that only b of $Z_1, Z_2, ..., Z_n$ will be 1 and others be 0. Then we can write

$$\|\nabla f_B(y) - \nabla f_B(x)\|^2 = \|\frac{1}{b}\sum_{i=1}^{n}(\nabla f_i(y) - \nabla f_i(x))Z_i\|^2 = \frac{1}{b^2}\|\sum_{i=1}^{n}(\nabla f_i(y) - \nabla f_i(x))Z_i\|^2$$

$$= \frac{1}{b^2}\left[\sum_{i=1}^{n}\|\nabla f_i(y) - \nabla f_i(x)\|^2 Z_i^2 + \sum_{i=1}^{n}\sum_{j\neq i}\langle f_i(y) - \nabla f_i(x), f_j(y) - \nabla f_j(x)\rangle Z_i Z_j\right]$$

Taking expectation on both sides, since

$$\mathbb{E}[Z_i^2] = \mathsf{Var}[Z_i] + \mathbb{E}[Z_i]^2 = \frac{b}{n} \quad , \quad \mathbb{E}[Z_i Z_j] = \frac{b(b-1)}{n(n-1)}$$

Properties of Mini-batch Function

We have

$$\begin{aligned}
&\|\nabla f_B(y) - \nabla f_B(x)\|^2 \\
&= \frac{1}{b^2}\left[\sum_{i=1}^{n} \|\nabla f_i(y) - \nabla f_i(x)\|^2 \frac{b}{n} + \sum_{i=1}^{n}\sum_{j \neq i} \langle f_i(y) - \nabla f_i(x), f_j(y) - \nabla f_j(x)\rangle \frac{b(b-1)}{n(n-1)}\right] \\
&= \frac{1}{b^2}\left[\left(\frac{b}{n} - \frac{b(b-1)}{n(n-1)}\right)\sum_{i=1}^{n} \|\nabla f_i(y) - \nabla f_i(x)\|^2 + \frac{b(b-1)}{n(n-1)} \|\sum_{i=1}^{n} \nabla f_i(y) - \nabla f_i(x)\|^2\right]
\end{aligned}$$

Since

$$\mathbb{E}[\|\nabla f_i(y) - \nabla f_i(x)\|^2] = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(y) - \nabla f_i(x)\|^2$$

$$\|\nabla f(y) - \nabla f(x)\|^2 = \|\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(y) - \nabla f_i(x)\|^2$$

Properties of Mini-batch Function

We have

$$
\begin{aligned}
&\|\nabla f_B(y) - \nabla f_B(x)\|^2 \\
&= \frac{1}{b^2}\left(\frac{b}{n} - \frac{b(b-1)}{n(n-1)}\right) n\mathbb{E}[\|\nabla f_i(y) - \nabla f_i(x)\|^2] + \frac{1}{b^2}\frac{b(b-1)}{n(n-1)} n^2 \|\nabla f(y) - \nabla f(x)\|^2 \\
&= \frac{n-b}{b(n-1)}\mathbb{E}[\|\nabla f_i(y) - \nabla f_i(x)\|^2] + \frac{n(b-1)}{b(n-1)}\|\nabla f(y) - \nabla f(x)\|^2 \\
&\leq 2\left(\frac{n-b}{b(n-1)}L_{\mathsf{max}} + \frac{n(b-1)}{b(n-1)}L_{\mathsf{avg}}\right)(f(y) - f(x) - \langle \nabla f(x), y - x\rangle)
\end{aligned}
$$

By Definition 3.1, since

$$\|\nabla f_B(y) - \nabla f_B(x)\|^2 \leq 2\mathcal{L}_b(f(y) - f(x) + \langle \nabla f(x), y - x\rangle)$$

We must have

$$\mathcal{L}_b \leq \frac{n(b-1)}{b(n-1)}L_{\mathsf{avg}} + \frac{n-b}{b(n-1)}L_{\mathsf{max}}$$

Properties of Mini-batch Function

Remark 3.5: By Lemma 3.4 we let

$$\sigma_b^* = \frac{n-b}{b(n-1)}\sigma_f^*$$

$$\mathcal{L}_b = \frac{n(b-1)}{b(n-1)} L_{\mathsf{avg}} + \frac{n-b}{b(n-1)} L_{\mathsf{max}}$$

When $b = 1$, we have $\sigma_b^* = \sigma_f^*$ and $\mathcal{L}_b = L_{\mathsf{max}}$, then mini-batch GD reduces to SGD.
When $b = n$, we have $\sigma_b^* = 0$ and $\mathcal{L}_b = L_{\mathsf{avg}}$, then mini-batch GD reduces to GD.
This shows mini-batch interpolates between single batch and full batch.

Convergence Analysis of Mini-Batch GD

The following Theorems are analogy to Theorem 2.1 and Theorem 2.3. Just consider each iteration calculates the gradient a mini-batch function f_B instead of the single function f_i, and replacing $L_{\max}$ and σ_f^* with $\mathcal{L}_b$ and σ_b^*.

Theorem 3.6: Let Assumptions (Sum of $L_{\max}$-smooth) and (Sum of Convex) hold. Consider a sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by Mini-batch GD, with the batch size b and the step size in each iteration satisfying $0 < t_k < \frac{1}{2\mathcal{L}_b}$, then

$$\mathbb{E}[f(\bar{x}^k) - \inf f] \leq \frac{\|x^0 - x^*\|^2}{2\sum_{i=0}^{k-1} t_i(1 - 2t_i\mathcal{L}_b)} + \frac{\sum_{i=0}^{k-1} t_i^2}{\sum_{i=0}^{k-1} t_i(1 - 2t_i\mathcal{L}_b)}\sigma_b^*$$

where $\bar{x}^k = \sum_{i=0}^{k-1} p_{i,k} x^k$, with $p_{i,k} = \frac{t_i(1-2t_i\mathcal{L}_b)}{\sum_{j=0}^{k-1} t_j(1-2t_j)\mathcal{L}_b}$.

Corollary 3.7: Consider the settings of Theorem 3.1. Let $t = \min\{\frac{\epsilon}{2(2\mathcal{L}_b\epsilon+\sigma_b^*)}, \frac{1}{2\mathcal{L}_b}\}$, then the iteration complexity of Mini-batch GD is $\Omega(\frac{1}{\epsilon^2})$.

Convergence Analysis of Mini-Batch GD

Theorem 3.8: Let Assumption (Sum of L_{max}) and (Sum of Convex) hold, and assume further that f is μ-strongly convex. Consider the $\{x^k\}_{k\in\mathbb{N}}$ sequence generated by the Mini-batch GD algorithm with batch size b and a constant step size satisfying $0 \leq t \leq \frac{1}{2\mathcal{L}_b}$. Then for $t \geq 0$,

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - t\mu)^k \|x^0 - x^*\|^2 + \frac{2t}{\mu}\sigma_b^*$$

Corollary 3.9: Consider the settings of Theorem 3.3. Let

$$t = \min\left\{\frac{\epsilon}{2}\frac{\mu}{2\sigma_b^*}, \frac{1}{2\mathcal{L}_b}\right\}$$

then

$$k \geq \max\left\{\frac{1}{\epsilon}\frac{4\sigma_b^*}{\mu^2}, \frac{2\mathcal{L}_b}{\mu}\right\}\log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon}\right) \quad \Longrightarrow \quad \|x^k - x^*\| \leq \epsilon$$

which means the iteration complexity of Mini-batch is $\Omega(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$.

Optimal Mini-batch Size

By Remark 3.5, we know that Mini-batch GD is a trade-off between GD and SGD. GD has low iteration complexity but high computation costs per iteration. SGD has low computation cost per iteration but high iteration complexity. By introducing Mini-batch GD, we can consider total cost as a function with respect to batch size, and find the optimal batch size to minimize the total cost such that it will be smaller than both GD and SGD.

Using the settings of Corollary 3.9, the number of iterations is

$$k \geq \max\left\{\frac{1}{\epsilon}\frac{4\sigma_b^*}{\mu^2}, \frac{2\mathcal{L}_b}{\mu}\right\}\log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon}\right)$$

and the computation cost per iteration is $O(bd)$, Thus the total cost is

$$T(b) \geq \max\left\{\frac{4d}{\epsilon\mu^2}b\sigma_b^*, \frac{2d}{\mu}b\mathcal{L}_b\right\}\log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon}\right)$$

Since

$$b\sigma_b^* = \frac{n-b}{(n-1)}\sigma_f^*, b\mathcal{L}_b = \frac{n(b-1)}{n-1}L_{\mathsf{avg}} + \frac{n-b}{n-1}L_{\mathsf{max}} = \frac{b(nL_{\mathsf{avg}} - L_{\mathsf{max}}) + n(L_{\mathsf{max}} - L_{\mathsf{avg}})}{n-1}$$

Optimal Mini-batch Size

We know that $b\sigma_b^*$ decreases with b and $b\mathcal{L}_b$ increases with b. Thus to minimize $\max\left\{\frac{4d}{\epsilon\mu^2}b\sigma_b^*, \frac{2d}{\mu}b\mathcal{L}_b\right\}$, we need to let

$$\frac{4d}{\epsilon\mu^2}b\sigma_b^* = \frac{2d}{\mu}b\mathcal{L}_b$$

By solving this we get the optimal batch size[8]:

$$b^* = \frac{\frac{2}{\epsilon\mu}n\sigma_f^* - n(L_{\mathsf{max}} - L_{\mathsf{avg}})}{nL_{\mathsf{avg}} - L_{\mathsf{max}} + \frac{2}{\epsilon\mu}\sigma_f^*}$$

[8] This result is exactly equation (38) in reference [6]

References

[1] Handbook of Convergence Theorems for (Stochastic) Gradient Methods. https://gowerrobert.github.io/pdf/M2_statistique_optimisation/grad_conv.pdf. (Important!)

[2] Iteration Complexity
https://www.cs.ubc.ca/~schmidtm/Courses/540-W18/L5.pdf

[3] SGD Convergence Rate
https://www.cs.ubc.ca/~schmidtm/Courses/540-W19/L11.pdf

[4] https://stanford.edu/~rezab/classes/cme323/S15/notes/lec11.pdf

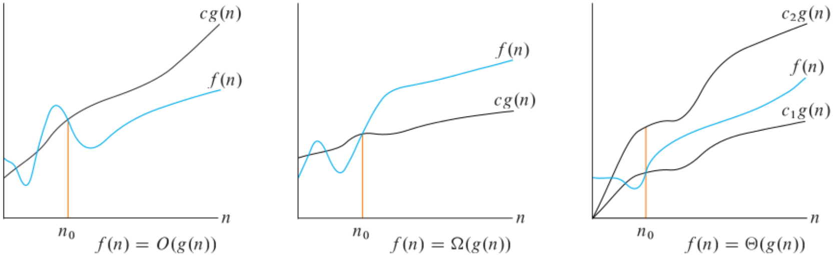
[5] Finite Population Sampling
http://www.et.bs.ehu.es/~etptupaf/nuevo/ficheros/stat4econ/muestreo.pdf

[6] R. M. Gower et al. "SGD: General Analysis and Improved Rates". International Conference on Machine Learning. 2019, pp. 5200–5209.
https://arxiv.org/pdf/1901.09401.pdf

Appendix 1: Asymptotic Notations: O, Ω, Θ

O, Ω, Θ are three common notations of asymptotic analysis for algorithms. Let c, c_1, c_2, n_0, n be positive real numbers, f and g be two positive functions, then

- $f(n) = O(g(n))$ if $\exists c, \exists n_0, \forall n > n_0 : 0 \leq f(n) \leq cg(n)$
- $f(n) = \Omega(g(n))$ if $\exists c, \exists n_0, \forall n > n_0 : 0 \leq cg(n) \leq f(n)$
- $f(n) = \Theta(g(n))$ if $\exists c_1, c_2, \exists n_0, \forall n > n_0 : 0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n)$



$f(n) = O(g(n))$ $f(n) = \Omega(g(n))$ $f(n) = \Theta(g(n))$

Thus, if $f(n) = \Theta(g(n))$, then $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

For more asymptotic notations, see Family of Bachmann-Landau notations in https://en.wikipedia.org/wiki/Big_O_notation.

Appendix 2: Harmonic Number

We call the series of all positive unit fractions $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, ...$ as Harmonic Series. The partial sum of the first n terms of Harmonic series $H_n = \sum_{k=1}^{n} \frac{1}{k}$ is called a Harmonic Number. [9]

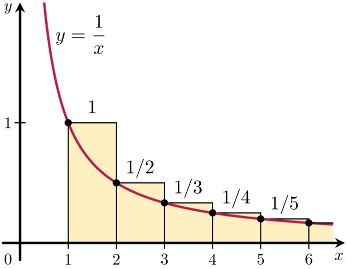
Lemma A.1: $H_n = \sum_{k=1}^{n} \frac{1}{k} = \Theta(\log n)$

Proof: Consider the figure on the right. Since the curve $y = \frac{1}{x}$ is below the upper boundary of the rectangles, we have

$$\int_1^{n+1} \frac{1}{x} dx < \sum_{k=1}^{n} \frac{1}{k}$$

Shifting each rectangle to the left by 1 unit, then $y = \frac{1}{x}$ will be above the upper boundary of the rectangles

$$\sum_{k=1}^{n} \frac{1}{k} < 1 + \int_1^{n} \frac{1}{x} dx$$



[9] https://en.wikipedia.org/wiki/Harmonic_series_(mathematics)

Appendix 2: Harmonic Number

Thus,

$$\int_1^{n+1} \frac{1}{x} dx < \sum_{k=1}^{n} \frac{1}{k} < 1 + \int_1^{n} \frac{1}{x} dx$$

$$\ln(n+1) < \sum_{k=1}^{n} \frac{1}{k} < 1 + \ln n$$

The notation $\Theta(\log n)$ here means there exists real numbers $k_1 > 0, k_2 > 0$ and integer n_0 such that for any $n > n_0$, we have $k_1 \log n < \sum_{k=1}^{n} \frac{1}{k} < k_2 \log n$ [10].

Let $n_0 = 3$, $k_1 = 1$, $k_2 = 2$. Then it is obvious that for any $n > 3$,

$$\ln n < \sum_{k=1}^{n} \frac{1}{k} < 2 \ln n$$

Since $\ln n = \log_e n$, for any base a, we have $\ln n = \log_a n / \log_a e = \ln a \log_a n$, which can be considered as $\log_a n$ times a constant $\ln a$. So it does not matter what the base is in $\Theta(\log n)$.

[10] https://en.wikipedia.org/wiki/Big_O_notation#Family_of_Bachmann-Landau_notations

Appendix 3: Lemma for Complexity

Lemma A.2: Consider the recurrence given by

$$\alpha_k \leq (1 - t\mu)^k \alpha_0 + At$$

where $\mu > 0$ and $A, C > 0$ are given constants and $t < \frac{1}{C}$. If

$$t = \min\left\{\frac{\epsilon}{2A}, \frac{1}{C}\right\}$$

Then,

$$k \geq \max\left\{\frac{1}{\epsilon}\frac{2A}{\mu}, \frac{C}{\mu}\right\} \log\left(\frac{2\alpha_0}{\epsilon}\right) \quad \Longrightarrow \quad \alpha_k \leq \epsilon$$

Proof: Since

$$t = \min\left\{\frac{\epsilon}{2A}, \frac{1}{C}\right\} \quad \Longleftrightarrow \quad \frac{1}{t} = \min\left\{\frac{2A}{\epsilon}, C\right\}$$

We have

$$\begin{aligned} k \geq \max\left\{\frac{1}{\epsilon}\frac{2A}{\mu}, \frac{C}{\mu}\right\} \log\left(\frac{2\alpha_0}{\epsilon}\right) &\Rightarrow k \geq \frac{1}{\mu} \max\left\{\frac{2A}{\epsilon}, C\right\} \log\left(\frac{2\alpha_0}{\epsilon}\right) \\ &\Rightarrow k \geq \frac{1}{t\mu} \log\left(\frac{2\alpha_0}{\epsilon}\right) \Rightarrow kt\mu \geq \log\left(\frac{2\alpha_0}{\epsilon}\right) \end{aligned}$$

Appendix 3: Lemma for Complexity

Since $\log(\frac{1}{\rho}) \geq 1 - \rho$ for $0 < \rho \leq 1$. Applying $\rho = 1 - t\mu$, we get $\log(\frac{1}{1-t\mu}) \geq t\mu$. Thus,

$$\begin{aligned} kt\mu \geq \log\left(\frac{2\alpha_0}{\epsilon}\right) &\Rightarrow k\log(\frac{1}{1-t\mu}) \geq \log\left(\frac{2\alpha_0}{\epsilon}\right) \\ &\Rightarrow \log\left(\frac{1}{1-t\mu}\right)^k \geq \log\left(\frac{2\alpha_0}{\epsilon}\right) \\ &\Rightarrow \left(\frac{1}{1-t\mu}\right)^k \geq \left(\frac{2\alpha_0}{\epsilon}\right) \ \Rightarrow \ \alpha_0(1-t\mu)^k \leq \frac{\epsilon}{2} \end{aligned}$$

Moreover, since

$$t = \min\left\{\frac{\epsilon}{2A}, \frac{1}{C}\right\} \quad \Longrightarrow \quad At = \min\left\{\frac{\epsilon}{2}, \frac{A}{C}\right\} \leq \frac{\epsilon}{2}$$

We have

$$\alpha_k = \alpha_0(1-t\mu)^k + At \leq \epsilon$$

Appendix 4: Finite Population Sampling

Lemma A.3: Suppose we have a finite population of size n. $y_1, y_2, ..., y_n$ are the elements of the population (they are considered as constants). Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ be the mean of the population, Var$(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ be the variance of the population. Let $B \subset \{1, 2, ..., n\}$ and $|B| = b < n$, and $\bar{Y} = \frac{1}{b}\sum_{i \in B} y_i$ be the estimator of $\bar{y}$ based on a sample of size b. Let Var$(\bar{Y}) = \frac{1}{\binom{n}{b}}\sum_{B}(\bar{Y} - \bar{y})^2$ be the variance of $\bar{Y}$, then

$$\text{Var}(\bar{Y}) = \frac{n-b}{b(n-1)}\text{Var}(y)$$

Proof: Let $Z_1, Z_2, ..., Z_n$ be random variables where each $Z_i \in \{0, 1\}$. $Z_i = 1$ means $i \in B$ and $Z_i = 0$ means $i \notin B$. We can write

$$\begin{aligned}\text{Var}(\bar{Y}) &= \text{Var}(\frac{y_1 Z_1 + y_2 Z_2 + ... + y_n Z_n}{b}) \\ &= \frac{1}{b^2}\left[\sum_{i=1}^{n} y_i^2 \text{Var}(Z_i) + \sum_{i=1}^{n}\sum_{j \neq i} y_i y_j \text{Cov}(Z_i, Z_j)\right]\end{aligned}$$

So we only need expressions for Var(Z_i) and Cov(Z_i, Z_j).

Appendix 4: Finite Population Sampling

We can consider B as selecting b objects from n **without replacement**. Thus the probability for one object to be chosen is,

$$P(Z_i = 1) = \frac{\binom{n-1}{b-1}}{\binom{n}{b}} = \frac{b}{n}$$

Here $\binom{n}{b} = \frac{n!}{(n-b)!b!}$. Thus each Z_i follows binary distribution, we have

$$\mathsf{E}[Z_i] = \frac{b}{n}, \quad \mathsf{Var}(Z_i) = \frac{b}{n}(1 - \frac{b}{n})$$

Since $\mathsf{Cov}(Z_i, Z_j) = \mathsf{E}[Z_i Z_j] - \mathsf{E}[Z_i]\mathsf{E}[Z_j]$. $Z_i Z_j$ is also a random variable from binary distribution where $Z_i Z_j = 1$ only when $Z_i = 1$ and $Z_j = 1$, which means two objects are chosen. Thus

$$\mathsf{E}[Z_i Z_j] = P(Z_i = 1, Z_j = 1) = \frac{\binom{n-2}{b-2}}{\binom{n}{b}} = \frac{b(b-1)}{n(n-1)}$$

$$\mathsf{Cov}(Z_i, Z_j) = \frac{b(b-1)}{n(n-1)} - (\frac{b}{n})^2 = -\frac{b(1-b/n)}{n(n-1)}$$

Appendix 4: Finite Population Sampling

Plugging in Var(Z_i) and Cov(Z_i, Z_j) we get

$$\begin{aligned}
\text{Var}(\bar{Y}) &= \frac{1}{b^2}\left[\sum_{i=1}^{n} y_i^2 \frac{b}{n}(1-\frac{b}{n}) - \sum_{i=1}^{n}\sum_{j\neq i} y_i y_j \frac{b(1-b/n)}{n(n-1)}\right] \\
&= \frac{1}{b^2}\frac{b}{n}(1-\frac{b}{n})\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} y_i y_j\right] \\
&= \frac{n-b}{bn^2}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} y_i y_j\right]
\end{aligned}$$

We can consider Var(y) as a special case of Var$(\bar{Y})$ when B has only one element, i.e., $b = 1$. Thus

$$\text{Var}(\bar{Y}) = \frac{n-b}{b}\frac{1}{n-1}\text{Var}(y) = \frac{n-b}{b(n-1)}\text{Var}(y)$$

This proof is from reference [5]. The only difference is that the Var(y) here is defined as population variance instead of quasi-variance.