# Optimization for Machine Learning

## Ch3: Stochastic Gradient Descent

**Dongyoung Lim**
**UNIST**

AI51101, IE55101

# Motivation

- Recall the population risk problem:

$$\min_{x \in \mathcal{X}} f(x) := \mathbb{E}_{Z \sim \mathcal{P}} \big[ L(x, Z) \big],$$

  where $Z \in \mathbb{R}^m$ and $\mathcal{P}$ is the (unknown) data distribution. Here $m$ denotes the number of data dimensions.

- Since $\mathcal{P}$ is unknown, we minimize the empirical risk:

$$\min_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} L(x, z_i),$$

  with a set of training samples $\{z_1, z_2, ..., z_n\}$ where $n$ is the number of samples.

- Gradient descent requires $\mathcal{O}(nm)$ computations *per iteration*, which becomes very expensive when both $n$ and $m$ are large.

# Motivation

- The key idea of Stochastic Gradient Descent (SGD) is to update parameters using a single randomly chosen sample:

$$x_{t+1} = x_t - \eta_t \, G(x_t, z_{i_t}),$$

where $i_t \in \{1, \ldots, n\}$ is chosen uniformly at random and $G$ is an unbiased estimator of the true gradient:

$$\mathbb{E}\big[G(x, Z)\big] = \nabla f(x), \quad Z \sim \mathcal{P}.$$

# Introduction

- Formally, SGD is set up as follows. At each iteration, we randomly draw an index $i_t$ uniformly from $\{1, \ldots, n\}$.
- A typical choice of the stochastic gradient is

$$G(x_t, z_{i_t}) = \nabla_x L(x_t, z_{i_t}),$$

  and the update is performed as

$$x_{t+1} = x_t - \eta_t \nabla_x L(x_t, z_{i_t}).$$

- This reduces the per-iteration computation from $\mathcal{O}(nm)$ to $\mathcal{O}(m)$.
- SGD converges to the optimum in expectation, but introduces variance in the optimization process (good or bad?). The smaller the batch size, the larger this variance, since it is inversely proportional to the number of samples used to compute each gradient.

# Introduction

- Brief history: the concept of Stochastic Gradient Descent has its roots in the "stochastic approximation" method proposed by Robbins and Monro in 1951.
- In this seminal work, they introduced a stochastic procedure to find the roots of equations, which later became the foundation of modern optimization algorithms for machine learning and AI.

# Stochastic Gradient Descent

- The SGD algorithm can be written as:

**Algorithm** SGD

1: **for** $t = 1, 2, \ldots, T$ **do**

2: Sample $i_t$ uniformly from $\{1, \ldots, n\}$.

3: Compute the stochastic gradient $G(x_t, z_{i_t}) = \nabla_x L(x_t, z_{i_t})$.

4: Update the parameter:

$$x_{t+1} = x_t - \eta_t G(x_t, z_{i_t}).$$

5: **end for**

## Problem Setup

- We aim to solve the population risk minimization problem

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{Z \sim \mathcal{P}}\big[L(x, Z)\big],$$

  where $Z \in \mathbb{R}^m$ follows an unknown distribution $\mathcal{P}$.

- Let $x^* \in \arg\min f(x)$ and define the initial distance

$$D := \|x_0 - x^*\|.$$

- We assume that
  - **Convexity:** $f$ is convex on $\mathbb{R}^d$.
  - **Unbiased stochastic gradients:** with $G(x_t, z_{i_t})$ computed from a single sample $z_{i_t}$,

$$\mathbb{E}[G(x, Z)] = \nabla f(x).$$

  - **Second-moment bound:**

$$\mathbb{E}\Big[\|G(x, Z)\|^2\Big] \le \sigma^2 \quad \text{for all } t.$$

# Unconstrained Case: Convex & second-moment bound

**Theorem (SGD Convergence)**

Consider stochastic gradient descent

$$x_{k+1} = x_k - \eta\, G(x_k, Z_k), \qquad k = 0, 1, \ldots, T-1.$$

With constant step size

$$\eta = \frac{D}{\sigma\sqrt{T}},$$

the averaged iterate

$$\bar{x}_T := \frac{1}{T}\sum_{k=0}^{T-1} x_k$$

satisfies

$$\mathbb{E}\big[f(\bar{x}_T)\big] - f(x^*) \le \frac{\sigma D}{\sqrt{T}}.$$

- Convergence rate: $\mathcal{O}\big(1/\sqrt{T}\big)$.

# Unconstrained Case: $\beta$-Smooth, Bounded, Nonconvex

- We study the problem

$$\min_{x \in \mathbb{R}^d} f(x),$$

where

- $f$ is (not necessarily convex but) $\beta$**-smooth**: $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$,
- $f$ is **bounded**: $|f(x)| \leq M$ for all $x$.

**Theorem (GD Stationarity Rate)**

Gradient descent with step size

$$\eta = \frac{1}{\beta}$$

satisfies

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \frac{4M\beta}{T}.$$

- The average squared gradient norm converges at rate $\mathcal{O}(1/T)$ even without convexity.

# Unconstrained Case: $\beta$-Smooth, Bounded, Nonconvex

- Same assumptions as before:
  - $f$ is $\beta$-smooth and $|f(x)| \leq M$,
  - $\mathbb{E}[G(x, Z)] = \nabla f(x)$,
  - $\mathbb{E}\|G(x, Z)\|^2 \leq \sigma^2$.

**Theorem (SGD Stationarity Rate, Constant $\eta$)**

With the constant stepsize

$$\eta = 2\sqrt{\frac{M}{\beta\,\sigma^2\,T}},$$

stochastic gradient descent satisfies

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right] \;\leq\; 2\sqrt{\frac{M\beta\sigma^2}{T}}.$$

- The expected average squared gradient norm decreases at rate $\mathcal{O}\left(1/\sqrt{T}\right)$.

# MINI-BATCH SGD

- In practice, we often use a variant of SGD called **mini-batch SGD**. In each iteration, it computes a stochastic gradient using a mini-batch of size $b$:

$$g_k := \frac{1}{b} \sum_{i=1}^{b} \nabla_x \ell\big(x_k, \mathbf{x}_{i_k^{(i)}}\big),$$

  where each index $i_k^{(i)}$ is sampled uniformly (typically without replacement) from $\{1, \dots, n\}$.
- This reduces the gradient variance by a factor of $b$, because under independent sampling

$$\mathrm{Var}\left( \frac{1}{b} \sum_{i=1}^{b} \nabla_x \ell(x_k, \mathbf{x}_{i_k^{(i)}}) \right) = \frac{1}{b^2} \sum_{i=1}^{b} \mathrm{Var}\left[ \nabla_x \ell(x_k, \mathbf{x}_{i_k^{(i)}}) \right] = \frac{\sigma^2}{b}.$$