Lecture 27 - 19th March

# Module C: Algorithms for Optimization

Recall that an optimization problem in standard form is given by

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \le 0, i \in [m] := \{1, 2, \ldots, m\}, \\ & h_j(x) = 0, j \in [p]. \end{aligned}$$

Most algorithms generate a sequence $x_0, x_1, x_2, \ldots$ by exploiting local information collected on the path.

$x_t \to x_{t+1}$

- Zeroth Order: Only $f(x_t), g_i(x_t), h_j(x_t)$ available.
- First Order: Gradients $\nabla f(x_t), \nabla g_i(x_t), \nabla h_j(x_t)$ are used. Heavily used in ML.
- Second Order: Hessian information is used. Eg: Newton's Method, etc.
- Distributed Algorithms
- Stochastic/Randomized Algorithms

## Measure of progress

Let $x^\star$ be the optimal solution. The iterative algorithms continue till any of the following error metrics is sufficiently small.

- $\texttt{err}_t := ||x_t - x^\star||$
- $\texttt{err}_t := f(x_t) - f(x^\star)$
- A solution $\bar{x}$ is $\epsilon$-optimal when

$$f(\bar{x}) \leq f(x^\star) + \epsilon.$$

  We often run the algorithm till $\texttt{err}_t$ is smaller than a sufficiently small $\epsilon > 0$.

- In presence of constaints, we define

  $\texttt{err}_t := \max(f(x_t) - f(x^\star), g_1(x_t), g_2(x_t), \ldots, g_m(x_t), |h_1(x_t)|, \ldots, |h_p(x_t)|).$

## First order methods: Gradient descent

Consider the unconstrained optimization problem: $\min_{x \in \mathbb{R}^n} f(x)$

> Gradient Descent (GD): $x_{t+1} = x_t - \eta_t \nabla f(x_t)$, $t \geq 0$ starting from an initial guess $x_0 \in \mathbb{R}^n$.

$x_1 = x_0 - \eta_0 \nabla f(x_0), \quad x_2 = x_1 - \eta_1 \nabla f(x_1) \ldots$

The stationarity condition satisfies $x^* = x^* - \eta_t \nabla f(x^*) \implies \nabla f(x^*) = 0$.

Convergence rate depends on choice of step size $\eta_t$ and characteristic of the function.

- Bounded Gradient: $||\nabla f(x)|| \leq G$ for all $x \in \mathbb{R}^n$.

- Smooth: A differentiable convex $f$ is $\beta$-smooth if for any $x, y$, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} ||y - x||^2. \quad = g_x(y) : \text{quadratic}$$

  We can obtain a quadratic upper bound on the function from local information.

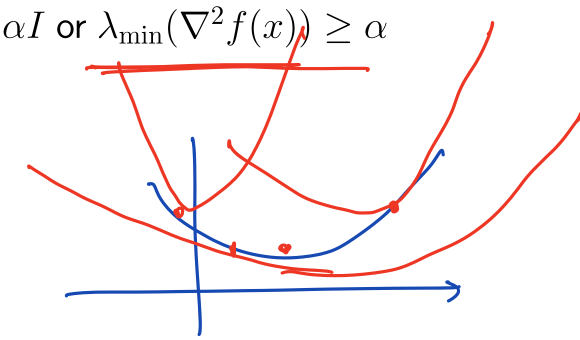
- Strongly Convex: A differentiable convex $f$ is $\alpha$-strongly convex if for any $x, y$, we have

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

  We can obtain a quadratic lower bound on the function from local information.

- If $f$ is twice differentiable, then
  - $f$ is $\beta$-smooth if and only if $\nabla^2 f(x) \preceq \beta I$ or $\lambda_{\max}(\nabla^2 f(x)) \leq \beta$ for all $x \in \mathbb{R}^n$.
  - $f$ is $\alpha$-strongly convex if and only if $\nabla^2 f(x) \succeq \alpha I$ or $\lambda_{\min}(\nabla^2 f(x)) \geq \alpha$ for all $x \in \mathbb{R}^n$.

- Determine $\beta$ and $\alpha$ for $f(x) = \frac{1}{2}||Ax - b||_2^2$.



$\nabla^2 f(x) = A^T A$, $\quad \beta = \lambda_{max}(A^T A)$

$\alpha = \lambda_{min}(A^T A)$

# Gradient Descent with Bounded Gradient Assumption

Let $x_0, x_1, \ldots, x_{T-1}$ be the iterates generated by the GD algorithm.

For any $t$, we define $\widehat{x}_t := \frac{1}{t}\sum_{i=0}^{t-1} x_i$. Let $x^\star$ be the optimal solution.

**Theorem 1: Convergence of Gradient Descent**

Let the function $f$ satisfy the $||\nabla f(x)|| \leq G$ for all $x \in \mathbb{R}^n$. Let $||x_0 - x^\star|| \leq D$. Then, for the choice of step size $\eta_t = \frac{D}{G\sqrt{T}}$, we have

$$f(\widehat{x}_T) - f(x^\star) \leq \frac{DG}{\sqrt{T}}. = \varepsilon \Rightarrow DG = \sqrt{T}\varepsilon \Rightarrow T = \frac{(DG)^2}{\varepsilon^2}$$

To find an $\epsilon$ optimal solution, choose $T \geq \left(\frac{DG}{\epsilon}\right)^2$ and $\eta = \frac{\epsilon}{G^2}$.

Possible Limitation: Need to know $G$ and $D$.

$x_{t+1} = x_t - \eta \nabla f(x_t)$

Proof: Define the following (potential) function:

$$\Phi_t := \frac{1}{2\eta}||x_t - x^\star||_2^2. \Rightarrow \phi_0 = \frac{D^2}{2\eta}$$

We show that $\Phi_t$ is decreasing in $t$. We compute $\Phi_{t+1} - \Phi_t$ as:

$$\phi_{t+1} - \phi_t = \frac{1}{2\eta}\left[ ||x_{t+1} - x^*||_2^2 - ||x_t - x^*||_2^2 \right] = \frac{1}{2\eta}\left[ ||x_{t+1} - x_t + x_t - x^*||_2^2 - ||x_t - x^*||_2^2 \right]$$

$$= \frac{1}{2\eta}\left[ ||x_{t+1} - x_t||_2^2 + 2\underbrace{\langle x_{t+1} - x_t, x_t - x^* \rangle}_{\text{inner product}} + \cancel{||x_t - x^*||_2^2} - \cancel{||x_t - x^*||_2^2} \right]$$

$$= \frac{1}{2\eta}\left[ \eta^2 ||\nabla f(x_t)||_2^2 + 2\langle -\eta \nabla f(x_t), x_t - x^* \rangle \right]$$

$$= \frac{\eta}{2}||\nabla f(x_t)||_2^2 - \langle \nabla f(x_t), x_t - x^* \rangle$$

$$\leq \frac{\eta}{2}G^2 - [f(x_t) - f(x^*)]$$

Recall that for a convex function,

$$f(x^*) \geq f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle$$

$$\Rightarrow \langle \nabla f(x_t), x_t - x^* \rangle \geq f(x_t) - f(x^*)$$

# Proof

Thus, we obtain

$$\Phi_{t+1} - \Phi_t \leq \frac{\eta}{2}G^2 - \left[f(x_t) - f(x^*)\right]$$

$$\Rightarrow f(x_t) - f(x^*) + \underbrace{\Phi_{t+1} - \Phi_t} \leq \frac{\eta}{2}G^2$$

adding the above equation from $t=0$ to $t=T-1$ gives us

$$\left[\begin{array}{l} \Phi_1 - \Phi_0 \\ \Phi_2 - \Phi_1 \\ \Phi_3 - \Phi_2 \\ \vdots \\ \Phi_T - \Phi_{T-1} \end{array}\right.$$

adding them leaves us with $\Phi_T - \Phi_0$

$$\sum_{t=0}^{T-1} f(x_t) - Tf(x^*) + \Phi_T - \Phi_0 \leq \frac{\eta T}{2}G^2$$

$$\Rightarrow \underbrace{\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*)}_{\to \text{LHS}} \leq \frac{\eta}{2}G^2 + \Phi_0 - \Phi_T \leq \frac{\eta}{2}G^2 + \frac{D^2}{2\eta T}$$

find $\eta$ to minimize

$$g(\eta) = \frac{\eta}{2}G^2 + \frac{D^2}{2\eta T}$$

$$g'(\eta) = \frac{G^2}{2} - \frac{D^2}{2\eta^2 T}$$

$$g''(\eta) > 0 \Rightarrow g \text{ is convex}$$

setting $g'(\eta) = 0$, we obtain

$$TG^2 = \frac{D^2}{\eta^2}$$

$$\Rightarrow \bar{\eta} = \frac{D}{G\sqrt{T}}$$

Since the function is convex,

$$f\Big(\underbrace{\frac{1}{T}\sum_{t=0}^{T-1} x_t}_{\hat{x}_T}\Big) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t)$$

$$\Rightarrow f(\hat{x}_T) - f(x^*) \leq \text{LHS} \leq \frac{\bar{\eta}}{2}G^2 + \frac{D^2}{2\bar{\eta}T}$$

$$= \frac{DG}{\sqrt{T}}.$$

## Gradient Descent with Smoothness Assumption

Recall that a differentiable convex $f$ is $\beta$-smooth if for any $x, y$, we have

$y \leftarrow x_{t+1}$, $x \leftarrow x_t$

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

**Theorem 2**

Let the function $f$ be $\beta$-smooth. Let $||x_0 - x^\star|| \leq D$. Then, for the choice of step size $\eta_t = \frac{1}{\beta}$, we have

$$f(x_T) - f(x^\star) \leq \frac{\beta||x_0 - x^\star||^2}{2T}. = \frac{\beta D^2}{2T}$$

Proof: Define the following (potential) function:

$$\Phi_t := t[f(x_t) - f(x^\star)] + \frac{\beta}{2}||x_t - x^\star||^2.$$

$$\phi_0 = \frac{\beta}{2}||x_0 - x^*||^2$$

We show that $\Phi_t$ is decreasing in $t$. We compute $\Phi_{t+1} - \Phi_t$ as:

If we can show that $\phi_T \leq \phi_0$

$$\Rightarrow T[f(x_T) - f(x^*)] + (\text{const}) \leq \frac{\beta}{2}||x_0 - x^*||_2^2$$

$$\Rightarrow f(x_T) - f(x^*) \leq \frac{\beta}{2T}||x_0 - x^*||_2^2.$$

Thus, it remains to show that $\phi_{t+1} \leq \phi_t \quad \forall t$.

$$\phi_{t+1} - \phi_t = \begin{cases} (t+1)[f(x_{t+1}) - f(x^*)] + \frac{\beta}{2}||x_{t+1} - x^*||^2 \\ - t\underbrace{[f(x_t) - f(x^*)]} - \frac{\beta}{2}||x_t - x^*||^2 \end{cases}$$

$$= \underline{(t+1)\left[f(x_{t+1}) - f(x^*)\right]} - \underline{(t+1)\left[f(x_t) - f(x^*)\right]} + f(x_t) - f(x^*)$$

$$+ \frac{\beta}{2}\left[\underbrace{\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2}\right]$$

**Proof**

---

$$\leq (t+1)\left[f(x_{t+1}) - f(x_t)\right] + \left[f(x_t) - \cancel{f(x^*)}\right]$$

(following the earlier proof)

$$+ \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2 - \cancel{\left[f(x_t) - f(x^*)\right]}$$

$$= (t+1)\left[\underline{f(x_{t+1}) - f(x_t)}\right] + \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

$$\leq (t+1)\left[\langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|_2^2\right] + \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

$$= (t+1)\left[-\frac{1}{\beta}\|\nabla f(x_t)\|_2^2 + \frac{\beta}{2}\frac{1}{\beta^2}\|\nabla f(x_t)\|_2^2\right] + \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

$$\frac{-1}{2\beta}\|\nabla f(x_t)\|_2^2.$$

$$= \|\nabla f(x_t)\|_2^2\left[-\frac{t+1}{\beta} + \frac{1}{2\beta}(t+1) + \frac{1}{2\beta}\right]$$

$$= \|\nabla f(x_t)\|_2^2\left[\frac{t+2-2t-2}{2\beta}\right]$$

$$= -\frac{t}{2\beta}\|\nabla f(x_t)\|_2^2 \leq 0.$$

$f(x) = \frac{1}{2}(x_1^2 + 100x_2^2)$

Lecture 28 : 20th March

$x_0 = (100, 100)$

$\beta = 100$

$\nabla f(x_t) = \begin{bmatrix} x_1 \\ 100x_2 \end{bmatrix}$

$x_1 = \begin{bmatrix} 100 \\ 100 \end{bmatrix} - \begin{bmatrix} 1 \\ 100 \end{bmatrix} = \begin{bmatrix} 99 \\ 0 \end{bmatrix}$

$\alpha = 1$

# Gradient Descent with Smoothness and Strong Convexity

$\kappa = 100$

$x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t), \quad = x_t - \begin{bmatrix} x_1/100 \\ x_2 \end{bmatrix}$

Recall that a differentiable convex $f$ is $\alpha$-strongly convex if for any $x, y$, we have

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

**Theorem 3**

Let the function $f$ be $\beta$-smooth and $\alpha$-strongly convex with $\alpha \leq \beta$. Define condition number $\kappa := \frac{\beta}{\alpha}$. Then, for the choice of step size $\eta_t = \frac{1}{\beta}$, we have

$$f(x_T) - f(x^\star) \leq e^{-\frac{T}{\kappa}}(f(x_0) - f(x^\star)).$$

Note: To obtain $\epsilon$-optimal solution, choose $T = \mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$.

Proof: Define the following (potential) function:

$$\Phi_t := (1+\gamma)^t[f(x_t) - f(x^\star)], \qquad \text{where } \boxed{\gamma = \frac{1}{\kappa - 1} = \frac{\alpha}{\beta - \alpha}}.$$

We need to show that $\Phi_{t+1} \leq \Phi_t$. $\Rightarrow \Phi_T \leq \Phi_0$

$\Rightarrow (1+\gamma)^T[f(x_T) - f(x^*)] \leq f(x_0) - f(x^*)$

$\Rightarrow f(x_T) - f(x^*) \leq (1+\gamma)^{-T}[f(x_0) - f(x^*)]$

It can be shown that

$(1+\gamma)^{-T} \leq e^{-T/\kappa}$.

we compute

$\Phi_{t+1} - \Phi_t$

$= (1+\gamma)^{t+1}[f(x_{t+1}) - f(x^*)]$

$\quad - (1+\gamma)^t[f(x_t) - f(x^*)]$

$\Rightarrow (1+\gamma)^{-t}[\Phi_{t+1} - \Phi_t]$

$= (1+\gamma)[f(x_{t+1}) - f(x^*)] - [f(x_t) - f(x^*)] + \gamma f(x_t) - \gamma f(x_t)$

$= (1+\gamma)[f(x_{t+1}) - f(x_t)] - (1+\gamma)f(x^*) + f(x^*) + \gamma f(x_t)$

$= (1+\gamma)[f(x_{t+1}) - f(x_t)] + \gamma[f(x_t) - f(x^*)]$

$$\leq (1+\gamma)\left(\frac{-1}{2\beta}\right)\|\nabla f(x_t)\|_2^2 + \gamma\left[f(x_t) - f(x^*)\right].$$

**Proof**

---

To simplify the second term, we will use the fact that $f$ is $\alpha$-strongly convex, which implies

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\alpha}{2}\|y-x\|_2^2 \quad \forall x, y$$

$$\Rightarrow f(x) - f(y) \leq \langle \nabla f(x), x-y\rangle - \frac{\alpha}{2}\|y-x\|_2^2.$$

If we choose $x \to x_t$, $y \to x^*$, we obtain

$$f(x_t) - f(x^*) \leq \langle \nabla f(x_t), x_t - x^*\rangle - \frac{\alpha}{2}\|x_t - x^*\|_2^2.$$

(PL-inequality)

$$\leq c_1^2 \|\nabla f(x_t)\|_2^2 = \frac{1}{2\alpha}\|\nabla f(x_t)\|_2^2.$$

Recall: $\|a-b\|_2^2 \geq 0 \iff a^T a \geq 2a^T b - b^T b$

$$a = c_1 \nabla f(x_t)$$

$$b = \sqrt{\frac{\alpha}{2}}\,(x_t - x^*)$$

$$2c_1\sqrt{\frac{\alpha}{2}} = 1 \Rightarrow c_1 = \frac{1}{\sqrt{2\alpha}}$$

Coming back:

$$(1+\gamma)^{-t}\left[\Phi_{t+1} - \Phi_t\right] \leq -\frac{(1+\gamma)}{2\beta}\|\nabla f(x_t)\|_2^2 + \frac{\gamma}{2\alpha}\|\nabla f(x_t)\|_2^2.$$

$$\frac{\gamma}{2\alpha} - \frac{1+\gamma}{2\beta} = 0 \qquad \leq 0$$

, Recall $\gamma = \frac{1}{\kappa - 1}$, $\kappa = \frac{\beta}{\alpha}$

$$\Rightarrow \Phi_{t+1} \leq \Phi_t \Rightarrow \Phi_T \leq \Phi_0 \quad \forall T \geq 0$$

# Summary of gradient descent convergence rates

Consider the unconstrained optimization problem: $\min_{x \in \mathbb{R}^n} f(x)$

Gradient Descent (GD): $x_{t+1} = x_t - \eta_t \nabla f(x_t), \quad t \geq 0$ starting from an initial guess $x_0 \in \mathbb{R}^n$.

**Theorem 4: GD Convergence rates**

Let $||x_0 - x^\star|| \leq D$.

- If $||\nabla f(x)|| \leq G$ for all $x \in \mathbb{R}^n$, then with $\eta_t = \frac{D}{G\sqrt{T}}$, $f(\widehat{x}_T) - f(x^\star) \leq \frac{DG}{\sqrt{T}}$. $\leq \varepsilon$
- If $f$ is $\beta$-smooth, for $\eta_t = \frac{1}{\beta}$, $f(x_T) - f(x^\star) \leq \frac{\beta||x_0 - x^\star||^2}{2T}$.
- If $f$ is $\beta$-smooth and $\alpha$-strongly convex, for $\eta_t = \frac{1}{\beta}$, $f(x_T) - f(x^\star) \leq e^{-\frac{T}{\kappa}}(f(x_0) - f(x^\star))$ where $\kappa := \frac{\beta}{\alpha}$ is the condition number.

To obtain an $\varepsilon$-optimal solution, we can choose $T$ as follows.

(1) $\frac{DG}{\sqrt{T}} \leq \varepsilon \Rightarrow T \geq \frac{D^2G^2}{\varepsilon^2}$

(2) $\frac{\beta D^2}{2T} \leq \varepsilon \Rightarrow T \geq \frac{\beta D^2}{2\varepsilon}$

(3) $e^{-T/\kappa} C \leq \varepsilon \Rightarrow e^{T/\kappa} \geq \frac{C}{\varepsilon} \Rightarrow \underline{T \geq \kappa \ln\left(\frac{C}{\varepsilon}\right)}$

## Gradient descent: Constrained Case

Consider the unconstrained optimization problem: $\min_{x \in X} f(x)$ where $X \subseteq \mathbb{R}^n$ is a convex feasibility set.

Projected Gradient Descent (PGD): $x_{t+1} = \Pi_X[x_t - \eta_t \nabla f(x_t)], \quad t \geq 0$ starting from an initial guess $x_0 \in \mathbb{R}^n$ where $\Pi_X(y)$ is the projection of $y$ on the set $X$.

**Theorem 5**

Let $||x_0 - x^\star|| \leq D$.

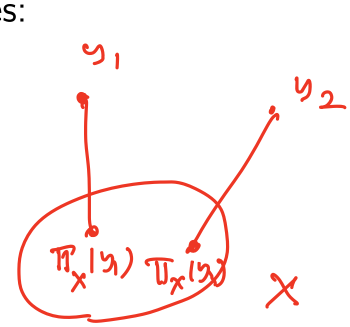
- If $||\nabla f(x)|| \leq G$ for all $x \in \mathbb{R}^n$, then with $\eta_t = \frac{D}{G\sqrt{T}}$, $f(\widehat{x}_T) - f(x^\star) \leq \frac{DG}{\sqrt{T}}$.
- If $f$ is $\beta$-smooth, for $\eta_t = \frac{1}{\beta}$, $f(x_T) - f(x^\star) \leq \frac{\beta||x_0 - x^\star||^2}{2T}$.
- If $f$ is $\beta$-smooth and $\alpha$-strongly convex, for $\eta_t = \frac{1}{\beta}$, $f(x_T) - f(x^\star) \leq e^{-\frac{T}{\kappa}}(f(0) - f(x^\star))$ where $\kappa := \frac{\beta}{\alpha}$ is the condition number.

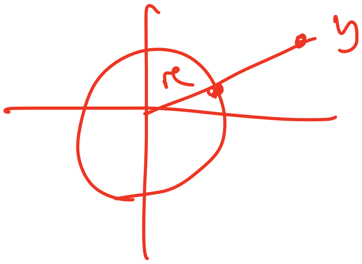
Note: Convergence rates remain unchanged.

Note: Projection itself is another optimization problem!

Non-expansive Property which preserves the convergence rates:

$$\boxed{||\Pi_X(y_1) - \Pi_X(y_2)|| \leq ||y_1 - y_2||.}$$



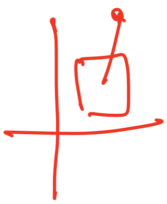



## When is Projection easy to find?

Note that $\Pi_X(y) = \text{argmin}_{x \in X} ||y - x||^2$. Find closed form expression of the projection for the following cases.

- $X_1 = \{x \in \mathbb{R}^n | ||x||_2 \le r\}$.

$$\Pi_{X_1}(y) = \frac{y}{||y||} \cdot r$$



- $X_2 = \{x \in \mathbb{R}^n | x_l \le x \le x_u\}$.

$$[\Pi_{X_2}(y)]_i = \begin{cases} y_i & \text{if } (x_l)_i \le y_i \le (x_u)_i \\ (x_u)_i & \text{if } y_i > (x_u)_i \\ (x_l)_i & \text{if } y_i < (x_l)_i \end{cases}$$

- $X_3 = \{x \in \mathbb{R}^n | Ax = b\}$.

Homework 2

- $X_4 = \{x \in \mathbb{R}^n | x \ge 0, \sum_{i=1}^n x_i \le 1\}$.

# Accelerated Gradient Descent

Start with $x_0 = y_0 = z_0 \in \mathbb{R}^n$. At every time-step $t$,

$$\begin{cases} y_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t) \\ z_{t+1} = z_t - \eta_t \nabla f(x_t) \\ x_{t+1} = (1-\tau_{t+1})y_{t+1} + \tau_{t+1} z_{t+1} \end{cases}$$

**Theorem 6**

Let $f$ be $\beta$-smooth, $\eta_t = \frac{t+1}{2\beta}$ and $\tau_t = \frac{2}{t+2}$. Then, we have

$$f(y_T) - f(x^\star) \leq \frac{2\beta||x_0 - x^*||^2}{T(T+1)}.$$

Proof: Define $\phi_t = t(t+1)(f(y_t) - f(x^*)) + 2\beta||z_t - x^*||^2$ and show that $\phi_{t+1} \leq \phi_t$.

# Accelerated Gradient Descent 2

Start with $x_0 = y_0$. At every state $t$,

$$y_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$$
$$x_{t+1} = (1 + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})y_{t+1} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}y_t$$

**Theorem 7**

Let $f$ be $\beta$-smooth, $\alpha$-strongly convex with $\kappa = \frac{\beta}{\alpha}$ and let $\gamma = \frac{1}{\sqrt{\kappa}-1}$. Then, we have

$$f(y_T) - f(x^\star) \leq (1+\gamma)^{-T}\left(\frac{\alpha+\beta}{2}||x_0 - x^*||^2\right).$$

Improvement upon the previous rate where we had $\gamma = \frac{1}{\kappa - 1}$.

## Further details

- AGD invented by Nesterov in a series of papers in the 80s and early 2000s, later popularized by ML researchers
- The convergence rates in the previous two theorems are the best possible ones. during the initial stage of the algorithm.
- Book by Nesterov: https://link.springer.com/book/10.1007/978-1-4419-8853-9
- https://francisbach.com/continuized-acceleration/
- https://www.nowpublishers.com/article/Details/OPT-036

---

Programming Tutorial

Let $x \in \mathbb{R}^d$, $c$: is a positive scalar, [write code treating c & d as variables, so that they can be varied)

$$f(x) = \left(c x_1^2 + \sum_{j=2}^{d} x_j^2\right) \times \frac{1}{2}$$

Let $x_0 = \begin{bmatrix} 100 \\ 100 \\ \vdots \\ 100 \end{bmatrix}$. Determine $x_T$ following GD & AGD with suitable step-size, and $T = 100$.

Plot: (1) $\log(f(x_t^{GD}))$ & $\log(f(x_t^{AGD}))$ vs. $t$ in the same figure

(2) for $d = 2$, plot $\left.\begin{matrix} x_{1,t}^{GD} \text{ vs. } x_{2,t}^{GD} \\ x_{1,t}^{AGD} \text{ vs. } x_{2,t}^{AGD} \end{matrix}\right\}$ in the same figure.

# Finite Sum Setting

- A large number of problems that arise in (supervised) ML can be written as

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x) = \frac{1}{N} \sum_{i=1}^{N} l(x, \xi_i).$$

ith data point

- Example: Regression/Least Squares, SVM, NN Training
- The above problem can also be viewed as *sample average approximation* of a stochastic optimization problem

$$f(x) = \mathbb{E}[l(x, \xi)]$$

involving uncertain parameter or random variable $\xi$.

- Challenge: $N$ (number of samples) or $n$ (dimension of decision variable) both may be large. Samples may be located in different servers.

$$\nabla f(x_t) = \frac{1}{N} \sum_{i=1}^{N} \nabla l(x_t, \xi_i)$$

Lecture-29: 26th March

## Gradient Descent vs. Stochastic Gradient Descent

Gradient Descent (GD) $x_{t+1} = x_t - \eta_t \nabla f(x_t) = x_t - \eta_t \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_t)$, $t \geq 0$ starting from an initial guess $x_0 \in \mathbb{R}^n$.

Each step requires $N$ gradient computations.

Stochastic Gradient Descent (SGD) At every time step $t$,

- Pick an index (sample) $i_t$ uniformly at random from the set $\{1, 2, \ldots, N\}$.
- Set $x_{t+1} = x_t - \eta_t \nabla f_{i_t}(x_t)$. $\rightarrow$ $\nabla \ell(x_t, \xi_{i_t})$

Each step requires $1$ gradient computation, which is a noisy version of the true gradient of the cost function at $x_t$.

$\nabla f_{i_t}(x_t)$ is a random variable, because index $i_t$ is a random variable.

$\mathbb{E}_{i_t}[\nabla f_{i_t}(x_t)] = \nabla f(x_t)$

$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$, in earlier notation, $f_i(x) = \ell(x, \xi_i)$

# Key result for SGD convergence

Under the following assumptions ← $\ell$ is convex in $x$ for fixed $\xi_i$

- Convexity: each $f_i$ is convex,
- Bounded variance: $\mathbb{E}[||\nabla f_{i_t}(x)||^2] \leq \sigma^2$ for some $\sigma$ for all $x$,
- Unbiased gradient estimate: $\mathbb{E}[\nabla f_{i_t}(x)] = \nabla f(x)$ for all $x$,

the solutions generated by SGD algorithm satisfies

$$\sum_{t=0}^{T-1} \eta_t [\mathbb{E}[f(x_t)] - f(x^\star)] \leq \frac{1}{2}||x_0 - x^\star||^2 + \frac{\sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2$$

$$\implies \mathbb{E}[f(\widehat{x}_T)] - f(x^\star) \leq \frac{||x_0 - x^\star||^2}{2 \sum_{t=0}^{T-1} \eta_t} + \frac{\sigma^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{t=0}^{T-1} \eta_t},$$

where $\widehat{x}_T = \frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \eta_t x_t$.

we need to choose $(\eta_t)_{t \geq 0}$ s.t.

$$\lim_{T\to\infty} \sum_{t=0}^{T-1} \eta_t = \infty,$$

$$\lim_{T\to\infty} \sum_{t=0}^{T-1} \eta_t^2 < \infty$$

- step-sizes that satisfy above conditions are called "square-summable"
- Robbins-Monro conditions

As before, we will

$$||x_{t+1} - x^*||_2^2 - ||x_t - x^*||_2^2$$

$$= ||x_{t+1} - x_t + x_t - x^*||_2^2 - ||x_t - x^*||_2^2$$

$$= ||x_{t+1} - x_t||_2^2 + 2\langle x_{t+1} - x_t, x_t - x^* \rangle$$

Recall: $x_{t+1} = x_t - \eta_t \nabla f_{i_t}(x_t)$, $i_t$ → random

$$= ||-\eta_t \nabla f_{i_t}(x_t)||_2^2 - 2\langle \nabla f_{i_t}(x_t), x_t - x^* \rangle$$

$$\leq \eta_t^2 \sigma^2 - 2\langle \nabla f_{i_t}(x_t), x_t - x^* \rangle$$

taking expectation on both sides,

## Proof Continues

$$\mathbb{E}\left[\|x_{t+1}-x^*\|_2^2 - \underbrace{\|x_t - x^*\|_2^2}\,\middle|\, x_t\right] \le \eta_t^2\sigma^2 - 2\left\langle \underline{\mathbb{E}[\nabla f_{j_t}(x_t)\,|\,x_t]},\, x_t - x^*\right\rangle$$

$$\Rightarrow \mathbb{E}\left[\|x_{t+1}-x^*\|_2^2 \,\middle|\, x_t\right] - \|x_t - x^*\|_2^2 \le \eta_t^2\sigma^2 - 2\langle \nabla f(x_t), x_t - x^*\rangle$$

$$\boxed{\begin{aligned} & f(x^*) \ge f(x_t) + \nabla f(x_t)(x^* - x_t) \\ \Rightarrow\ & \nabla f(x_t)(x_t - x^*) \ge f(x_t) - f(x^*) \\ \Rightarrow\ & -\left[\nabla f(x_t)(x_t - x^*)\right] \le f(x^*) - f(x_t) \end{aligned}}$$

$$\Rightarrow \underline{\mathbb{E}\left[\|x_{t+1}-x^*\|_2^2 \,\middle|\, x_t\right]} - \underline{\|x_t - x^*\|_2^2} \le \eta_t^2\sigma^2 + 2\left[f(x^*) - f(x_t)\right]$$

we can add both LHS & RHS for $t=0$ to $t=T$

$$\underline{\mathbb{E}\left[\|x_{t+1}-x^*\|_2^2\right]} - \|x_0 - x^*\|_2^2 \le \underline{\sigma^2 \sum_{t=0}^{T} \eta_t^2} + 2T \quad f(x^*)$$

$$- 2\sum_{t=0}^{T} f(x_t)$$

$$\Rightarrow 2\left[\sum_{t=0}^{T} f(x_t) - T f(x^*)\right] \le \sigma^2 \sum_{t=0}^{T} \eta_t^2 + \|x_0 - x^*\|_2^2 \quad - (*)$$

$$\le \underline{\sigma^2 \sum_{t=0}^{T} \eta_t^2 + \|x_0 - x^*\|_2^2}$$

we divide $2\sum_{t=0}^{T} \eta_t$ on both sides to obtain

$$\underbrace{\frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T} f(x_t) - \frac{T f(x^*)}{\sum_{t=0}^{T-1} \eta_t}}_{\text{VI}_2} \leq \underline{RHS}$$

$$\underline{f(\hat{x}_T) - f(x^*)}$$

using the definition of convex functions

$$f\left(\sum_{i=1}^{K} \lambda_i x_i\right) \leq \sum_{i=1}^{K} \lambda_i f(x_i)$$

for $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$

---

What if we choose a constant step-size?

$$\frac{\|x_0 - x^\star\|^2}{2\sum_{t=0}^{T-1} \eta_t} + \frac{\sigma^2 \sum_{t=0}^{T-1} \eta_t^2}{2\sum_{t=0}^{T-1} \eta_t}$$

$\sigma^2 T \eta^2$

$2T\eta$ $\quad$ $2T\eta$

Suppose $\eta_t = \eta$,

$$\underline{RHS}: \underbrace{\frac{\|x_0 - x^*\|^2}{2\eta T}}_{\text{decreasing in } T} + \frac{\sigma^2 \eta}{2}$$

does not depend on $T$

$$\mathbb{E}[f(\hat{x}_T)] - f(x^*) \leq \underline{\underline{\frac{\sigma^2 \eta}{2}}} \text{ as } T \to \infty$$

# Choice of stepsize

Constant step-size will not give us convergence. For convergence, we need to choose step sizes that are diminishing and square-summable, i.e.,

$$\lim_{T\to\infty} \sum_{t=0}^{T-1} \eta_t = \infty, \qquad \lim_{T\to\infty} \sum_{t=0}^{T-1} \eta_t^2 < \infty.$$

- If $\eta_t := \frac{1}{c\sqrt{t+1}}$, then $\mathbb{E}[f(\widehat{x}_T)] - f(x^\star) \leq \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$. This rate does not improve when the function is smooth.
- When the function is smooth, then for $\eta_t := \eta$ chosen appropriately, then R.H.S. will be of order $\mathcal{O}\left(\frac{1}{\eta T}\right) + \mathcal{O}(\eta)$.

## Analysis for Smooth and Strongly Convex Functions

When the function $f$ is $\beta$-smooth and $\alpha$-strongly convex, we have the following guarantees for SGD after $T$ iterations.

- If $\eta_t := \frac{1}{ct}$ for a suitable constant $c$, then error bound is $\mathcal{O}\left(\frac{\log T}{T}\right)$. Can be improved to $\mathcal{O}\left(\frac{1}{T}\right)$.
- If $\eta_t := \eta$, then error bound

$$\mathbb{E}[||x_T - x^\star||^2] \leq (1-\eta\alpha)^T ||x_0 - x^\star||^2 + \frac{\eta\beta\sigma^2}{2\alpha}.$$

With constant step-size $\eta < \frac{1}{\alpha}$, convergence is quick to a neighborhood of the optimal solution.

# Extension: Mini-Batch

- at any given time $t$, pick a set of indices $I_t \subseteq \{1, 2 \cdots N\}$ uniformly at random such that $|I_t| = b$

When $b = 1 \Rightarrow$ SGD
$b = N \Rightarrow$ GD

typically, choose $b \ll N$.

$\rightarrow$ $$x_{t+1} = x_t - \eta_t \cdot \underline{\frac{1}{b} \sum_{j \in I_t} \nabla f_j(x_t)}$$

$\rightarrow$ Convergence rate established on $\mathbb{E}[f(\hat{x}_T)]$ or $\mathbb{E}[\|x_T - x^*\|_2^2]$

remain unchanged, but the variance reduces by a factor $b$.

## Extension: Stochastic Averaging

- at time 0, define $g^0 = \frac{1}{N}\sum_{i=1}^{N} f_i(x_0)$, $g_i^0 = f_i(x_0)$
- at time $t$,
  - pick index $i_t$ at random
  - $g_i^t = \begin{cases} g_i^{t-1} & \text{if } i \neq i_t \\ \nabla f_{i_t}(x_t) & \text{if } i = i_t \end{cases}$
  - $x_{t+1} = x_t - \eta_t \frac{1}{N}\sum_{i=1}^{N} g_i^t$
- This scheme enjoys considerable advantages compared to SGD.

# Further Reading

SAG: Schmidt, Mark, Nicolas Le Roux, and Francis Bach. "Minimizing finite sums with the stochastic average gradient." Mathematical Programming 162 (2017): 83-112.

SAGA: Defazio, Aaron, Francis Bach, and Simon Lacoste-Julien. "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives." Advances in neural information processing systems 27 (2014).

Recent Review: Gower, Robert M., Mark Schmidt, Francis Bach, and Peter Richtárik. "Variance-reduced methods for machine learning." Proceedings of the IEEE 108, no. 11 (2020): 1968-1983.

Allen-Zhu, Zeyuan. "Katyusha: The First Direct Acceleration of Stochastic Gradient Methods." Journal of Machine Learning Research 18 (2018): 1-51.

Varre, Aditya, and Nicolas Flammarion. "Accelerated SGD for non-strongly-convex least squares." In Conference on Learning Theory, pp. 2062-2126. PMLR, 2022.

Hanzely, Filip, Konstantin Mishchenko, and Peter Richtárik. "SEGA: Variance reduction via gradient sketching." Advances in Neural Information Processing Systems 31 (2018).

## Extension: Adaptive Step-sizes

$$[x_{t+1}]_i = [x_t]_i - \eta_t \frac{[\nabla f(x_t)]_i}{|\nabla f(x_t)|_i}$$ : gradient normalization

$\Downarrow$

helps in problems that are badly conditioned & to avoid saddle points.

AdaGrad Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." Journal of machine learning research 12, no. 7 (2011).

Adam Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).