Module B: Algorithms for Optimization

Recall that an optimization problem in standard form is given by

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

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

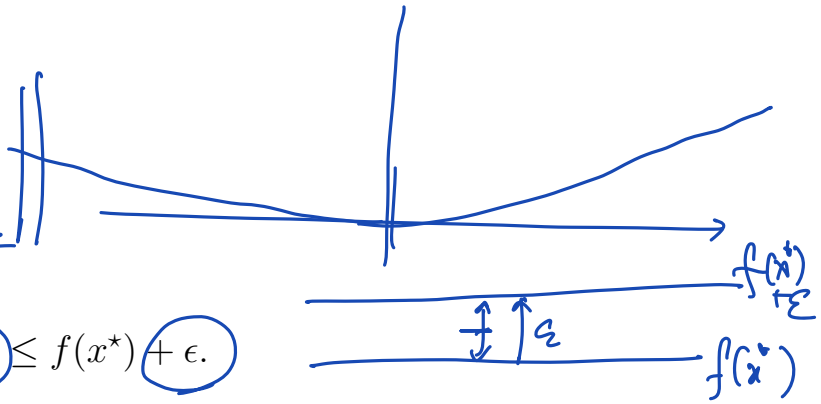
- Zeroth Order: Only $f(x_t), g_i(x_t), h_j(x_t)$ available.
- First Order: Gradients $\nabla f(x_t), \nabla g_i(x_t), \nabla h_j(x_t)$ are used. Heavily used in ML.
- Second Order: Hessian information is used. Eg: Newton's Method, etc.
- Distributed Algorithms
- Stochastic/Randomized Algorithms

Measure of progress

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

- $\texttt{err}_t := ||x_t - x^\star||_2^2$
- $\texttt{err}_t := f(x_t) - f(x^\star)$
- A solution $\bar{x}$ is ϵ-optimal when

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



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

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

First order methods: Gradient descent

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

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

Convergence rate depends on choice of step size η_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 β-smooth if for any x, y, we have

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

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

- Strongly Convex: A differentiable convex f is α-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 β-smooth $\Leftrightarrow \nabla^2 f(x) \preceq \beta I \ \forall x \in \mathbb{R}^n$

$\Leftrightarrow \lambda_{max}(\nabla^2 f(x)) \leq \beta \ \forall x \in \mathbb{R}^n$

f is α-strongly convex $\Leftrightarrow \nabla^2 f(x) \succeq \alpha I \quad \forall x \in \mathbb{R}^n$

$\Leftrightarrow \lambda_{min}[\nabla^2 f(x)] \geq \alpha \quad \forall x \in \mathbb{R}^n$

Ex: $f(x) = ||Ax - b||_2^2$

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

$\|\nabla f(x)\| \le G \quad \forall x \in \mathbb{R}^n$.

Gradient Descent with Bounded Gradient Assumption

First pick ϵ. ✓
Then set T, ✓
then find η, ✓
then run GD.

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.

take avg. of soln's.

$\widehat{x}_T$ has ϵ-optimality.

Theorem 1: Convergence of Gradient Descent

Let the function f satisfy the bounded gradient property. Let $\|x_0 - x^\star\| \le D$. Then, for the choice of step size $\eta_t = \frac{D}{G\sqrt{T}}$, we have

$$f(\widehat{x}_T) - f(x^\star) \le \frac{DG}{\sqrt{T}}.$$

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

$\frac{DG}{\sqrt{T}} \le \epsilon$

$\Rightarrow T \ge \left(\frac{DG}{\epsilon}\right)^2$

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

Possible Limitation: Need to know D and G.

10^{-3}, $\frac{1}{\epsilon^2} \simeq 10^6$

Proof: Define the following (potential) function:

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

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

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

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

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

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

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

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

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

Proof

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

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

$\Rightarrow \phi_{t+1} - \phi_t + f(x_t) - f(x^*) \leq \frac{\eta}{2} G^2, \quad t = 0, 1, 2, \ldots, T$

$\sum_{t=0}^{T-1} \left[\phi_{t+1} - \phi_t + f(x_t) - f(x^*) \right] \leq \frac{\eta T}{2} G^2$

$\Rightarrow \phi_T - \phi_0 + \sum_{t=0}^{T-1} f(x_t) - T f(x^*) \leq \frac{\eta}{2} T G^2$

$\Rightarrow \left[\frac{1}{T} \sum_{t=0}^{T-1} f(x_t) - f(x^*) \right] + \frac{1}{T} \phi_T \leq \frac{\eta}{2} G^2 + \frac{\phi_0}{T}$

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

find η which minimizes R.H.S.

$\frac{1}{2} G^2 = \frac{D^2}{2T} \cdot \frac{1}{\eta^2} \Rightarrow \eta^* = \frac{D}{G\sqrt{T}}$

$\Rightarrow \frac{1}{T} \sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \frac{G^2}{2} \cdot \frac{D}{G\sqrt{T}} + \frac{D^2}{2T} \cdot \frac{G\sqrt{T}}{D}$

$f(\hat{x}_T) - f(x^*) = \frac{GD}{2\sqrt{T}} + \frac{GD}{2\sqrt{T}} = \frac{GD}{\sqrt{T}}$

From convexity of f, $f\left(\frac{1}{T}\sum_{t=0}^{T-1} x_t\right) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t)$

Proof Continues

Finally $f(\hat{x}_T) - f(x^*) \leq \frac{GD}{\sqrt{T}}$

Gradient Descent with Smoothness Assumption

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

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

Theorem 2

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

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

Proof: Define the following (potential) function:

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

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

If we require ε-optimal soln, then $\frac{\beta ||x_0 - x^\star||^2}{2T} \leq \varepsilon$

$\Rightarrow T \geq \frac{\beta ||x_0 - x^\star||^2}{2\varepsilon} \approx O\left(\frac{1}{\varepsilon}\right)$

$\Phi_0 = \frac{\beta}{2} ||x_0 - x^\star||$

$\Phi_T = T\left[f(x_T) - f(x^\star)\right]$ (+ const)

If $\Phi_{t+1} \leq \Phi_t \; \forall t$, then $\Phi_T \leq \Phi_0$

$$\Rightarrow \boxed{f(x_T) - f(x^\star) \leq \frac{\beta ||x_0 - x^\star||}{2T}}$$

$$\Phi_{t+1} - \Phi_t = (t+1)\left[f(x_{t+1}) - f(x^\star)\right] + \frac{\beta}{2}||x_{t+1} - x^\star||^2 - t\left[f(x_t) - f(x^\star)\right] + \frac{\beta}{2}||x_t - x^\star||^2$$

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

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

Proof

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

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

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

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

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

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

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

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

$\Rightarrow \Phi_{t+1} \leq \Phi_t \Rightarrow \Phi_T \leq \Phi_0$ which proves the theorem.

Gradient Descent with Smoothness and Strong Convexity

Recall that a differentiable convex f is α-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 β-smooth and α-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 ϵ-optimal solution, choose $T = \mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$.

Proof: Define the following (potential) function:

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

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

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

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

HW $\leq e^{-T/\kappa} [f(x_0) - f(x^*)]$

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

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

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

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

$+ \gamma f(x_t) \qquad - \gamma f(x_t)$

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

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

$$\leq (1+\gamma)\left(\frac{-1}{2\beta}\|\nabla f(x_t)\|_2^2\right) + \frac{\gamma}{2\alpha}\|\nabla f(x_t)\|_2^2$$

1st term: use smoothness

Proof

Since f is β-smooth, $f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|_2^2$

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

$$= \langle \nabla f(x_t), -\frac{1}{\beta}\nabla f(x_t) \rangle + \frac{\beta}{2}\|-\frac{1}{\beta}\nabla f(x_t)\|_2^2$$

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

GD: $x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$

2nd term: use strong convexity.

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

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

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

Note: $\|a-b\|_2^2 = a^Ta + b^Tb - 2a^Tb \geq 0$

$$\Rightarrow a^Ta \geq 2a^Tb - b^Tb$$

let $b = \sqrt{\frac{\alpha}{2}}(x_t - x^*)$, $a = \frac{1}{\sqrt{2\alpha}}\nabla f(x_t)$

$f(x_t) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x_t)\|^2$ is called PL-inequality.

Strongly convex functions satisfy this.

Some non-convex functions also satisfy this.

Recall: $\gamma = \frac{\alpha}{\beta - \alpha} \Rightarrow 1+\gamma = \frac{\beta}{\beta-\alpha} \Rightarrow \frac{1+\gamma}{2\beta} = \frac{1}{2(\beta-\alpha)}$

Proof Continues

---

we have

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

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

$$= 0 .$$

$\Rightarrow \phi_{t+1} \leq \phi_t \quad \forall t.$

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 $\eta_t = \frac{D}{G\sqrt{T}}$, $f(\widehat{x}_T) - f(x^\star) \leq \frac{DG}{\sqrt{T}}$.
- If f is β-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 β-smooth and α-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.

$T \gtrsim \frac{C}{\varepsilon^2}$

$T \gtrsim \frac{C}{\varepsilon}$

If κ is large, $e^{-T/\kappa}$ decreases slowly $\Rightarrow$ ill-conditioned problems

If $f(x) = x^T x$, $\kappa = 1$ since $\nabla^2 f(x) = I$

$f(x) = 10^6 x_1^2 + \sum_{j=2}^{n} x_j^2$, $\kappa = 10^6$

for a desired ε,

$e^{-T/\kappa}(f(x_0) - f(x^\star)) \leq \varepsilon$

$\Rightarrow -\frac{T}{\kappa} + \log(\text{const}) \leq \log(\varepsilon)$

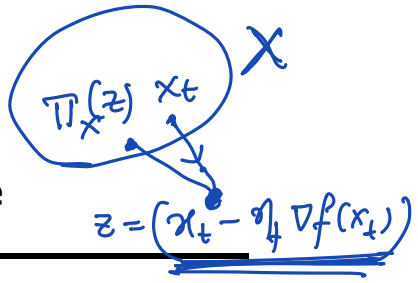
$\Rightarrow \frac{T}{\kappa} \gtrsim -\log(\varepsilon) + \text{const.}$

$\log(f(x_t) - f(x^\star))$

t

$\Rightarrow T \gtrsim \kappa \left[\log\left(\frac{1}{\varepsilon}\right) + \text{const}\right]$

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

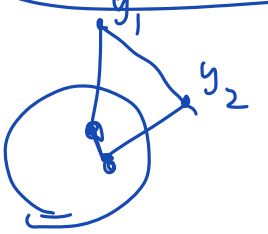
- If $||\nabla f(x)|| \leq G$ for all $x \in \mathbb{R}^n$, then with $\eta_t = \frac{D}{G\sqrt{T}}$, $f(\widehat{x}_T) - f(x^\star) \leq \frac{DG}{\sqrt{T}}$.
- If f is β-smooth, for $\eta_t = \frac{1}{\beta}$, $f(x_T) - f(x^\star) \leq \frac{\beta ||x_0 - x^\star||^2}{2T}$.
- If f is β-smooth and α-strongly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

$$||\Pi_X(y_1) - \Pi_X(y_2)|| \leq ||y_1 - y_2||. \quad (*)$$



$\frac{y}{\|y\|} \cdot r$

$\operatorname{argmin}_{x \in X_1} \|x-y\|_2^2$

to cross check, one should verify $\Pi_X(y) \in X$

When is Projection easy to find?

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

$y \in \mathbb{R}^n$

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

$$\Pi_{X_1}(y) = \frac{r\,y}{\|y\|_2}$$

$$\|\Pi_{X_1}(y)\|_2 = \left\| \frac{r\,y}{\|y\|_2} \right\|_2 = \frac{r}{\|y\|_2} \cdot \|y\|_2 = r$$

$\Rightarrow \Pi_{X_1}(y) \in X_1$

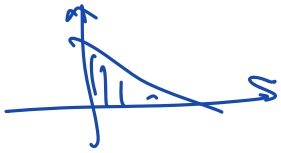
$x_l, x_u \in \mathbb{R}^n$

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

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

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

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



$$\Pi_{X_4}(y) = \begin{cases} y & \text{if } y \in X_4 \\ \max(y - \mu^*, 0) & \text{where } \mu^* \in \mathbb{R} \text{ is the solution of} \end{cases}$$

$$\sum_{i=1}^n \max(y_i - \mu, 0) = 1$$

HW

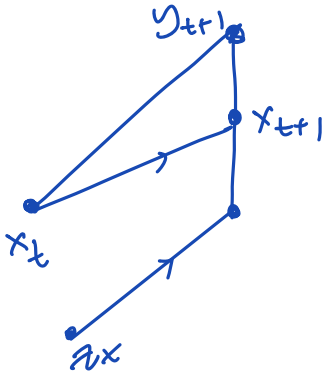
where $[\Pi_{X_4}(y)]_i = \max(y_i - \mu^*, 0)$

$(\Pi_{X_4}(y))_i$

Accelerated Gradient Descent

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

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



Theorem 6

Let f be β-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 $\phi_{t+1} \leq \phi_t$.

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

$\phi_T = T(T+1)[f(y_T) - f(x^*)] + \text{const}$.

If $\phi_{t+1} \leq \phi_t \; \forall t$, then $\phi_T \leq \phi_0$

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

$\phi_{t+1} \leq \phi_t$ $\updownarrow$ HW

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

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

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.
- 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

Finite Sum Setting

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

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

- 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 ξ.

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

Gradient Descent vs. Stochastic Gradient Descent

Gradient Descent (GD) $x_{t+1} = x_t - \eta_t \nabla f(x_t) = x_t - \eta_t \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_t), \quad t \geq 0$ 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 $x_{t+1} = x_t - \eta_t \nabla f_{i_t}(x_t)$.

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

$f(x) = \|Ax-b\|_2^2 = \sum_{i=1}^{N} (a_i^T x - b_i)^2$

$f_i(x) = (a_i^T x - b_i)^2$

Key result for SGD convergence

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

Under the following assumptions

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

$\nabla f_{i_t}(x) = \nabla f(x) + \varepsilon_{i_t}$

the solutions generated by SGD algorithm satisfies

$x_{t+1} - x_t = -\eta_t \nabla f_{i_t}(x_t)$

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

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

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

we compute

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

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

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

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

taking expectation on both sides:

$$\Delta\phi_t = \mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right] - \mathbb{E}\left[\|x_t - x^*\|_2^2\right]$$

$$= \eta_t^2 \mathbb{E}\left[\|\nabla f_{i_t}(x_t)\|_2^2\right] - 2\eta_t \mathbb{E}\left[\nabla f_{i_t}(x_t)^T (x_t - x^*)\right]$$

$\mathbb{E}_{i_t|x_t} = \mathbb{E}_{i_t}$

Proof Continues

Now:

$$\Delta\phi_t \le \eta_t^2\sigma^2 - 2\eta_t \left(\mathbb{E}_{i_t|x_t}\left[\nabla f_{i_t}(x_t)\right]\right)^T (x_t - x^*)$$

$$= \eta_t^2\sigma^2 - 2\eta_t \nabla f(x_t)^T (x_t - x^*)$$

$$\le \eta_t^2\sigma^2 - 2\eta_t \left[f(x_t) - f(x^*)\right]$$

Since function f is convex,
$f(x^*) \ge f(x_t) + \nabla f(x_t)^T (x^* - x_t)$
$\Rightarrow \nabla f(x_t)^T (x_t - x^*) \ge f(x_t) - f(x^*)$

$$\Rightarrow 2\eta_t \left[f(x_t) - f(x^*)\right] \le \eta_t^2\sigma^2 - \Delta\phi_t$$

$$= -\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right] + \mathbb{E}\left[\|x_t - x^*\|_2^2\right] + \eta_t^2\sigma^2$$

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

$$\le \sigma^2 \sum_{t=0}^{T-1} \eta_t^2 + \|x_0 - x^*\|_2^2$$

$$\Rightarrow \frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \eta_t f(x_t) - f(x^*) \le \frac{\sigma^2}{2} \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} + \frac{\|x_0 - x^*\|_2^2}{2\sum_{t=0}^{T-1} \eta_t}$$

$$f(\hat{x}_T) \le$$

$$\Rightarrow f(\hat{x}_T) - f(x^*) \le \left(\quad \right)$$

Proof Continues

Proof Continues

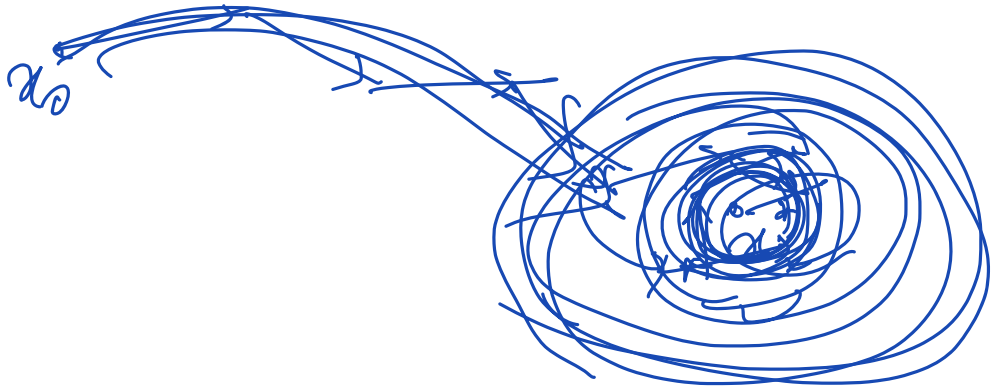
Choice of stepsize

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 $\mathcal{O}\left(\frac{1}{\eta T}\right) + \mathcal{O}(\eta)$.

Recall that for vanilla GD under bounded gradient assumption, we had $f(\widehat{x}_T) - f(x^*) \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$



Analysis for Smooth and Strongly Convex Functions

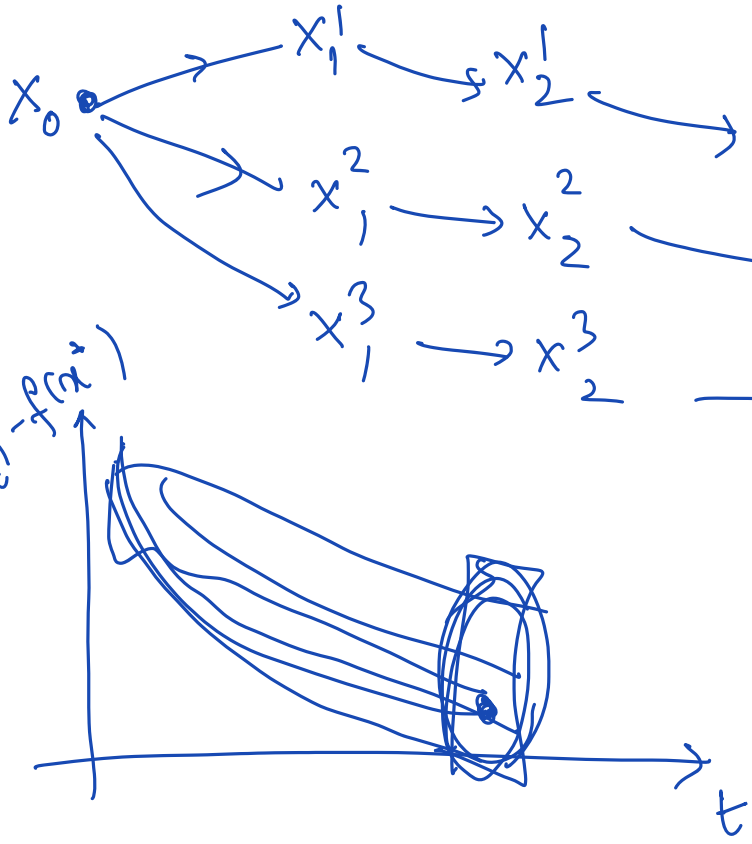
When the function f is β-smooth and α-strongly convex, we have the following guarantees for SGD after T iterations.

- 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



b : batch-size

at every t,

- pick $\{i_{t,1}, i_{t,2}, \dots, i_{t,b}\}$ uniformly at random
- $x_{t+1} = x_t - \eta_t \left[\frac{1}{b} \sum_{j=1}^{b} \nabla f_{t,j}(x_t) \right]$

When $b=1$: SGD

$b=N$: GD

The variance of error is smaller by a factor of b compared to vanilla SGD

Extension: Stochastic Averaging

$$x_{t+1} = x_t - \eta_t g_t, \qquad g_t = \frac{1}{N}\sum_{i=1}^{N} g_t^i,$$

$$g_t^i = \begin{cases} g_{t-1}^i & \text{if } i \neq i_t \\ \nabla f_{i_t}(x_t) & \text{if } i = i_t \end{cases}$$

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

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).