Optimization Algorithms for Machine/Deep Learning

Deterministic optimization methods

$minf(x), \quad x \in X$

f: differentiable

$||\triangledown f(x) - \triangledown f(y)|| \leq L||x-y||, \;\; \forall x, y \in X$

e.g. $f(x) = \frac{1}{2}X^{\mathrm{T}}AX - b^{\mathrm{T}}x$

$\quad\quad \triangledown f(x) = Ax - b$

$\quad\quad ||\triangledown f(x) - \triangledown f(y)|| = ||A(x-y)|| \leq ||A||||x-y||, \;\; L = ||A||$

Lemma $\;\; f(x) \leq f(y) + < \triangledown f(y), x-y > + \frac{1}{2}||x-y||^2$

Proof: $\;\;$ Let $\phi(t) = f(y + t(x-y))$

$\quad\quad \triangledown \phi(t) = \triangledown f(y + t(x-y))^{\mathrm{T}}(x-y)$

$\quad\quad \phi(1) - \phi(0) = \int_0^1 \triangledown \phi(t) dt$

$\quad\quad f(x) - f(y) = \int_0^1 \triangledown f(y + t(x-y))^{\mathrm{T}}(x-y)dt$

$f(x) - f(y) - < \triangledown f(y), x-y >= \int_0^1 \triangledown f(y + t(x-y))^{\mathrm{T}}(x-y)dt - \int_0^1 \triangledown f(y)^{\mathrm{T}}(x-y)dt$

$\quad\quad = \int_0^1 (\triangledown f(y + t(x-y)) - \triangledown f(y))^{\mathrm{T}}(x-y)dt$

$\quad\quad \leq \int_0^1 ||\triangledown f(y + t(x-y))||||x-y||dt$ (Cauchy inequality)

$\quad\quad \leq tL||x-y||^2 dt$

$\quad\quad = \frac{L}{2}||x-y||^2$

$x_{t+1} = argmin_{x \in X} Y_t < \triangledown f(x), x > + \frac{1}{2}||x - x_t||^2$

Optimality condition for the above subproblem

- $< \gamma_t \triangledown f(x_t) + x_{t+1} - x_t, x - x_{t+1} > \geq 0, \;\; \forall x \in X$ (OPT1)
- $\gamma < \triangledown f(x_t), x_{t+1} - x > \leq \frac{1}{2}||x - x_t||^2 - \frac{1}{2}||x - x_{t+1}|| - \frac{1}{2}||x_t - x_{t+1}||^2$ (OPT2)

Observation

$f(x_{t+1}) \leq f(x_t)$ if $\gamma_t \leq \frac{2}{L}$

Fix $x = x_t$ in OPT1, $< \triangledown f(x_t), x_{t+1} - x_t > \leq -\frac{1}{\gamma_t}||x_{t+1} - x_t||^2$

Also, $f(x_{t+1}) \leq f(x_t) + < \triangledown f(x_t), x_{t+1} - x_t > + \frac{L}{2}||x_{t+1} - x_t||^2$

$$\leq f(x_t) - (\frac{1}{\gamma_t} - \frac{L}{2})||x_{t+1} - x_t||^2$$

$$\leq f(x_t)$$

$$f(x_{t+1}) \leq f(x_t) + < \nabla f(x_t), x_{t+1} - x_t > + \frac{1}{2}||x_{t+1} - x||^2$$

$$= f(x_t) + < \nabla f(x_t), x - x_t > + < \nabla f(x_t), x_{t+1} - x > + \frac{L}{2}||x_{t+1} - x_t||^2$$

(Strong) Convexity OPT2

$$\leq f(x) + \frac{1}{2\gamma_t}[||x - x_t||^2 - ||x - x_{t+1}||^2] - \frac{1}{2}(\frac{1}{\gamma_t} - L)||x_{t+1} - x_t||^2$$

$$\leq f(x) + \frac{1}{2\gamma_t}[||x - x_t||^2 - ||x - x_{t+1}||^2]$$

If $\gamma_t = \gamma \leq \frac{1}{L}$, $t \leq 1, 2, \ldots$, then $\sum_{t=1}^{k}[f(x_{t+1}) - f(x)] \leq \frac{1}{2\gamma}[||x - x_t||^2 - ||x - x_{t+1}||^2]$

Notice $f(x_t) \geq f(x_{k+1})$, $\forall t \leq k+1$

$$\sum_{t=1}^{k}[f(x_{t+1}) - f(x)] \geq k[f(x_{k+1}) - f(x)]$$

Then $f(x_{k+1}) - f(x) \leq \frac{1}{2\gamma k}||x - x_1||^2$

$$\gamma = \frac{1}{L} \Rightarrow f(x_{k+1}) - f(x^*) \leq \frac{L}{2k}||x^* - x_1||^2$$

Strong Convexity

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

e.g. $f(x) = \frac{1}{2}x^{\mathrm{T}}Ax + b^{\mathrm{T}}x$, $\mu = \lambda_{min}(A)$

$$f(x_{t+1}) \leq f(x) - \frac{\mu}{2}||x - x_t||^2 + \frac{1}{2\gamma}[||x - x_t||^2 - ||x - x_{t+1}||^2]$$

$$f(x+1) - f(x^*) + \frac{1}{2\gamma}||x^* - x_{t+1}||^2 \leq \frac{1}{2}(\frac{1}{\gamma} - \mu)||x^* - x_t||^2$$

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

If $r = \frac{1}{L}$, then $||x_{t+1} - x^*|| \leq (1 - \frac{\mu}{L})||x_t - x^*||^2$

$$||x_{k+1} - x^*|| \leq (1 - \frac{\mu}{L})^k||x_t - x^*||^2$$

If we want to have $||x_{k+1} - x^*|| \leq \varepsilon$

It suffices to have $(1 - \frac{\mu}{L})^k||x_1 - x^*||^2 \leq \varepsilon$

$$(1 - \frac{\mu}{L})^k \leq \frac{\varepsilon}{||x_1 - x^*||^2}$$

$$k \cdot log(1 - \frac{\mu}{L}) \leq log\frac{\varepsilon}{||x_i - x^*||^2}$$

$$k \cdot (-log(1 - \frac{\mu}{L})) \geq log\frac{||x_i - x^*||^2}{\varepsilon}$$

$$k \geq \frac{1}{-log(1 - \frac{\mu}{L})}log\frac{||x_1 - x^*||^2}{\varepsilon} \Leftarrow l \geq \frac{L}{\mu}log\frac{||x_i - x^*||^2}{\varepsilon}$$

ε: conditional number

$\nabla f(x)$

$E[G(x_t, \xi_t)] = \nabla f(x_t)$

Define $\delta_t = \nabla f(x_t) - G(x_t, \xi_t)$

- $E[\delta_t] = 0$

  δ_t independent of x_t

- $E[||\delta_t||^2] \leq \sigma^2$

$x_{t+1} = argmin_{x \in X} \gamma_t < G(x_t, \xi_t), x > + \frac{1}{2}||x - x_t||^2$

$\gamma_t < G(x_t, \xi_t), x_{t+1} - x > \leq \frac{1}{2}[||x - x_t||^2 - ||x - x_{t+1}||^2 - ||x_t - x_{t+1}||^2]$ (OPT2')

`todo:`

Will be inplemented later or you can pull requests my Github Repo

Comments

1. $f(x) = E_\xi[F(x, \xi)]$, ξ is continuous random variable, SGD nearly optimal
2. $f(x) = \frac{1}{N}\sum_{t=1}^N f_i(x)$, using randomized incremental gradient method, we can improve the speed of convergence in terms of the dependence on ε. But the convergence depends on N.

- Deep Learning
- Burer-Monteiro Law Rank

  decomposition

  $X = LU, \ \ L \in \mathbb{R}^{m \times r}, U \in \mathbb{R}^{r \times n}$

  $min_{L,U}||X - LU||^2$

Nonconvex Optimization

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

f is smooth but not necessarily convex

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

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

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^{\mathrm{T}}(x_{t+1} - x_t) + \frac{L}{2}||x_{t+1} - x_t||^2 \\ &= f(x_1) - \gamma_t||\nabla f(x_t)||^2 + \frac{L\gamma_t^2}{2}||\nabla f(x_t)||^2 \\ &= f(x_t) - \gamma_t(1 - \frac{L\gamma_t}{2})||\nabla f(x_t)||^2 \end{aligned}$$

$\gamma_t(1 - \frac{L\gamma_t}{2})||\nabla f(x_t)||^2 \leq f(x_t) - f(x_{t+1})$

$\sum_{t=1}^k \gamma_t(1 - \frac{L\gamma_t}{2})||\nabla f(x_t)||^2 \leq f(x_1) - f(x_{t+1}) \leq f(x_1) - f^*$

Output $\overline{x_k}$ s.t. $||\nabla f(\overline{x_k})|| = min_{t=1,...,k}||\nabla f(\overline{x_t})||$

$0 \leq \gamma_t \leq \frac{2}{L}$

$\sum_{t=1}^{k} \gamma_t (1 - \frac{L\gamma_t}{2}) ||\triangledown f(x_t)||^2 \geq ||\triangledown f(\overline{x_k})||^2 \sum_{t=1}^{k} r_t (1 - \frac{Lr_t}{2})$

`todo:`

Will be inplemented later or you can pull requests my [Github Repo]

$\sum d_i = n, d_i >= 1$