## Lecture 21 - Catalyst

Tuesday, April 4, 2023 2:27 PM

today: catalyst → accelerate
non-convex opt.
submodular opt.

### catalyst algorithm

[Lin, Mairal & Harchaoui NeurIPS 2015]

"meta algorithm": outer loop which uses a linearly convergent alg. inner loop to get overall acceleration (!?)

main idea: use the accelerated proximal point alg.

with approximation inner loop of prox operator

proximal point alg.: is proximal gradient with $f=0$

$$\boxed{w_{t+1} = \text{prox}_{\gamma}^{\Omega}(w_t)}$$

[to solve $\min_w \Omega(w)$]

**catalyst alg.:** (for $\mu$-strongly $F(w)$)

repeat:

$$w_{t+1} \approx \underbrace{\text{argmin}_w\, F(w) + \frac{1}{2\gamma}\|w - z_t\|_2^2}_{\triangleq\, G_t(w)} \quad \text{s.t. } G_t(w_{t+1}) - \min_w G_t(w) \leq \varepsilon_t$$

to be specified

$\triangleq \text{prox}_{\gamma}^{F(\cdot)}(z_t)$

($\gamma$ is algorithmic parameter)

use inner loop optimization with warm start [e.g. SAGA, AFW, etc.]

$$z_{t+1} = w_{t+1} + \beta_{t+1}(w_{t+1} - w_t)$$

[accelerated Nesterov trick piece]

"extrapolation" / "momentum"

let $q \triangleq \dfrac{\mu}{\mu + \frac{1}{\gamma}}$ ; $\beta_{t+1}$ is found using fancy equations so that everything works

- solve for $\alpha_{t+1}$ in eq.: $\alpha_{t+1}^2 = (1-\alpha_{t+1})\alpha_t^2 + q\,\alpha_{t+1}$

(pick $\alpha_{t+1} \in ]0,1[$)

$$\beta_{t+1} \triangleq \frac{\alpha_t(1-\alpha_t)}{\alpha_t^2 + \alpha_{t+1}}$$

catalyst trick: use $\gamma$ & $\varepsilon_t$ s.t. overall # of inner loop calls gives an overall acceleration

with clever analysis of warm starting

### acceleration results

strong convexity of $G_t(w)$

- if inner loop alg. has convergence $\exp(-\frac{\tilde{\mu}}{L} t)$ $\quad \tilde{\mu} \geq \mu + \frac{1}{\gamma}$

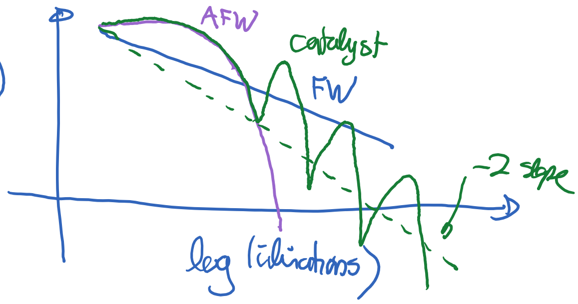
then with correct constants for $\gamma$ & $\varepsilon_t$

($\mu$-strongly convex $F$) linear rate $\rho = \frac{1}{\kappa}$ $\xrightarrow[\text{with catalyst}]{\text{becomes}}$ $\approx \frac{1}{\sqrt{\kappa}}$ for catalyst

($F$ convex case) $O(\frac{1}{t})$ on $F$ $\xrightarrow{\text{becomes}}$ $O(\frac{1}{t^2})$

result: we can get (theory) accelerated SAGA, SVRG, AFW etc...

issue: catalyst is not adaptive to local strong convexity
and finicky for choice of $\gamma$, $\varepsilon_t$, $\mu$ etc..




## non-convex optimization

recall: FW with line search on $f$ non-convex $\min_{s \leq t} g(w_s) \leq O(\frac{1}{\sqrt{t}})$

FW gap

convex $\quad \mathbb{E} f(w_t) - f^* \leq \varepsilon$

GD $\rightarrow \frac{1}{\varepsilon}$ $\qquad$ Nesterov $\rightarrow \frac{1}{\sqrt{\varepsilon}}$

non-convex: $\mathbb{E} \|\nabla f(w_t)\|_2^2 \leq \varepsilon$

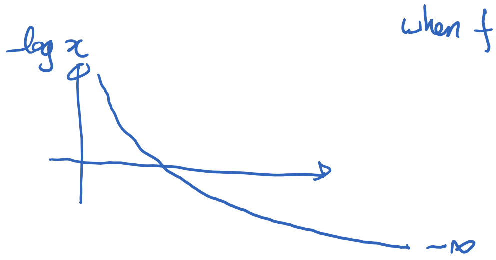
btw: if $f$ is $\mu$-strongly convex

$\Rightarrow f(w_t) - f^* \leq \frac{1}{2\mu} \|\nabla f(w_t)\|_2^2$

note: $\|\nabla f(w_t)\|$ small $\not\Rightarrow f(w_t) - f^*$ is small
when $f$ is not strongly convex

$-\log x$, $\varphi$,

when $f$ is not strongly convex




$*$ can get a $O(\frac{1}{\varepsilon})$ rate for gradient descent:

$$f(w) \leq f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + \frac{L}{2}\|w - w_t\|^2 \quad \forall w$$

$$w_{t+1} = w_t - \frac{1}{L}\nabla f(w_t) \qquad \text{(L-smoothness of } f \text{ but no need of convexity)}$$

$$\Rightarrow f(w_{t+1}) \leq f(w_t) - \frac{1}{2L}\|\nabla f(w_t)\|_2^2$$

If $f^*$ is finite

$$\Rightarrow f(w_{t+1}) - f^* \leq f(w_t) - f^* - \frac{1}{2L}\|\nabla f(w_t)\|_2^2$$

$$\leq f(w_{t-1}) - f^* - \frac{1}{2L}\left(\|\nabla f(w_t)\|^2 + \|\nabla f(w_{t-1})\|^2\right)$$

$$\leq \cdots f(w_0) - f^* - \frac{1}{2L}\left(\sum_{s=0}^{t}\|\nabla f(w_s)\|_2^2\right)$$

$$\Rightarrow \underbrace{\sum_{s}\|\nabla f(w_s)\|^2}_{} \leq 2L\left(f(w_0) - f^*\right)$$

$$\Rightarrow t \cdot \min_{s \leq t}\|\nabla f(w_s)\|_2^2 \leq \qquad \Rightarrow \boxed{\min_{s \leq t}\|\nabla f(w_s)\|^2 \leq \frac{2L}{t}\left(f(w_0) - f^*\right)} \quad //$$

NeurIPS 2016 tutorial "Large-Scale Optimization: Beyond Stochastic Gradient Descent and Convexity"
[Suvri Sra slides]

## Faster nonconvex optimization via VR

(Reddi, Hefny, Sra, Poczos, Smola, 2016; Reddi et al., 2016)

| Algorithm | Nonconvex (Lipschitz smooth) |
|---|---|
| SGD | $O(\frac{1}{\epsilon^2})$ |
| GD | $O(\frac{n}{\epsilon})$ |
| SVRG | $O(n + \frac{n^{2/3}}{\epsilon})$ |
| SAGA | $O(n + \frac{n^{2/3}}{\epsilon})$ |
| MSVRG | $O\left(\min\left(\frac{1}{\epsilon^2}, \frac{n^{2/3}}{\epsilon}\right)\right)$ |

$\mathbb{E}[\|\nabla g(\theta_t)\|^2] \leq \epsilon$

**Remarks**

New results for convex case too; additional nonconvex results
For related results, see also *(Allen-Zhu, Hazan, 2016)*

20

## Linear rates for nonconvex problems

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

The **Polyak-Łojasiewicz (PL)** class of functions

$$g(\theta) - g(\theta^*) \leq \frac{1}{2\mu}\|\nabla g(\theta)\|^2$$

*(Polyak, 1963); (Łojasiewicz, 1963)*

## Linear rates for nonconvex problems

$$g(\theta) - g(\theta^*) \leq \frac{1}{2\mu}\|\nabla g(\theta)\|^2 \quad \Big| \quad \mathbb{E}[g(\theta_t) - g^*] \leq \epsilon$$

| Algorithm | Nonconvex | Nonconvex-PL |
|---|---|---|
| SGD | $O\left(\frac{1}{\epsilon^2}\right)$ | $O\left(\frac{1}{\epsilon^2}\right)$ |
| GD | $O\left(\frac{n}{\epsilon}\right)$ | $O\left(\frac{n}{2\mu}\log\frac{1}{\epsilon}\right)$ |
| SVRG | $O\left(n + \frac{n^{2/3}}{\epsilon}\right)$ | $O\left(\left(n + \frac{n^{2/3}}{2\mu}\right)\log\frac{1}{\epsilon}\right)$ |
| SAGA | $O\left(n + \frac{n^{2/3}}{\epsilon}\right)$ | $O\left(\left(n + \frac{n^{2/3}}{2\mu}\right)\log\frac{1}{\epsilon}\right)$ |
| MSVRG | $O\left(\min\left(\frac{1}{\epsilon^2}\right), \frac{n^{2/3}}{\epsilon}\right)$ | — |

Variant of nc-SVRG attains this fast convergence!

*(Reddi, Hefny, Sra, Poczos, Smola, 2016; Reddi et al., 2016)* 22

15h35

## Submodular optimization

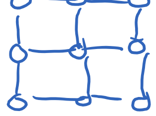
submodularity is an analog of convexity / concavity for tractability of set functions
(combinatorial opt.)

$F: 2^V \to \mathbb{R}$

$V = \{1, \dots, d\}$ is "ground set"

$2^V = \{ V \to \{0,1\}\}$ = set of all subsets of $V$

concrete example:




Ising model $E(y) = \sum_i \theta_i y_i - \sum_i \sum_{j \in \text{Neighbor}(i)} \theta_{ij} y_i y_j$

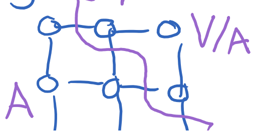
$A_y = \{ i : y_i = 1\}$

$F(A_y) = E(y)$

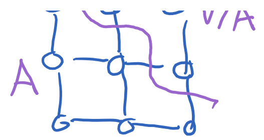
when $\theta_{ij} \geq 0$ "attractive potential" $\Rightarrow E(y)$ is submodular

MRF here is "associative Markov network" (AMN)

$\rightarrow$ can minimize $E(y)$ (or $F(A_y)$) by using "graph cut alg."




equivalent with min cost network flow problem




network flow problem

F is submodular $\Leftrightarrow$ $F(A)+F(B) \geq F(A\cap B) + F(A \cup B)$ $\forall A,B \subseteq V$

$\Leftrightarrow$ function $A \mapsto F(A \cup \{k\}) - F(A)$ is non-increasing for all $k$

i.e. $F(A \cup \{k\}) - F(A) \leq F(B \cup \{k\}) - F(B)$ if $B \subseteq A$

(a bit like concavity)

"diminishing return property"

$\Rightarrow$ intuitively, that greedy alg. are not "too bad" for maximization

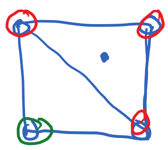
* $F(A) \triangleq g(|A|)$ cardinality — if $g$ is concave $\Rightarrow$ then F is submodular

* link with convexity $\rightarrow$ Lovasz extension (cf. fct. extension) $A \subseteq V$

* embed sets as corners of hypercube in dimension $d$ $\quad V(A) = \mathbb{1}_A \in \{0,1\}^d$

Lovasz extension $f$ extends $F(\cdot)$ from corners to entire hypercube using convex interpolation

(piecewise linear fct. on $[0,1]^d$)



$f(w) = F(A)$ when $w = V(A)$

let's say $w = \sum_i \alpha_i v_i$, $v_i = V(A_i)$ $\Rightarrow$ $f(w) = \sum_i \alpha_i \underbrace{F(A_i)}_{f(V(A_i))}$

$$\boxed{F \text{ is submodular} \Leftrightarrow \text{Lovasz extension } f \text{ is convex}}$$

(it turns out)

* can write $f(w) = \max\limits_{s \in B(F)} \langle s, w \rangle$ $\leftarrow$ this can be computed efficiently using sorting + greedy alg

$B(F)$: "base polytope"

(LMO over $B(F)$ is efficient)

$\min\limits_{A \subseteq V} F(A) = \min\limits_{w \in [0,1]^d} \underbrace{\left(\max\limits_{s \in B(F)} \langle s, w \rangle\right)}_{f(w)}$ $\rightarrow$ use projected subgradient method

$\partial f(w_t) = \arg\max\limits_{s \in B(F)} \langle s, w_t \rangle$

* with $\ell_2$-regularization, use duality to get a smooth obj. $\quad \min\limits_{s \in B(F)} \frac{1}{2}\|s\|^2$

$\rightarrow$ use "min-norm pt." alg. variant of FCFW alg. $\quad$ (+) SOTA for submodular opt.?

variant of FCFW alg.

submodular opt. ?