Lecture 21 - catalyst

Thursday, April 1, 2021 13:28

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. in inner loop to get overall acceleration (!)

main idea : use the accelerated proximal point algorithm with approximation (inner loop) of prox operator

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

$$w_{t+1} = \text{prox}_{\gamma}^{\Omega}(w_t) \qquad \text{(to solve } \min_w \Omega(w)\text{)}$$

catalyst alg. (for μ-strongly $F(w)$)

let $q \triangleq \frac{\mu}{\mu + \frac{1}{\gamma}}$ (γ is algorithmic parameter)

repeat:

$$w_{t+1} \approx \underset{w}{\text{argmin}} \; \underbrace{F(w) + \frac{1}{2\gamma}\|w - z_t\|_2^2}_{\triangleq G_t(w)} \quad \text{s.t. } G_t(w_{t+1}) - \min_w G_t(w) \le \varepsilon_t \text{ (to be specified)}$$

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

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

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

"extrapolation" / "momentum" [accelerated Nesterov trick piece]

β_{t+1} is found using fancy equations so that everything works

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

$$\beta_{t+1} \triangleq \frac{\alpha_t(1-\alpha_t)}{\alpha_t^2 + \alpha_{t+1}} \qquad \text{(pick } \alpha_{t+1} \in]0,1[\text{)}$$

catalyst trick: use γ & ε_t
s.t. overall # of inner loop calls gives an overall acceleration

with clever analysis of warm starting

acceleration results:

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

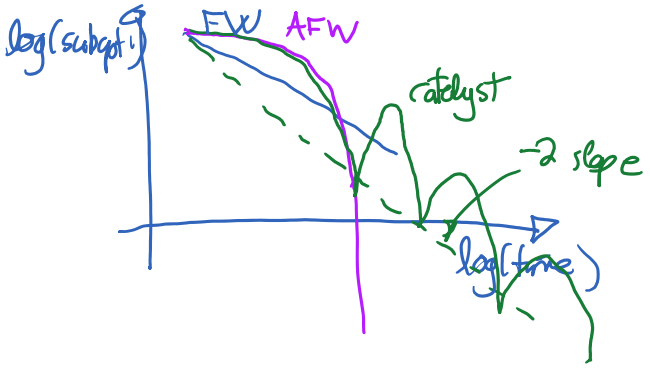
then with correct constants for γ & ε_t

(μ-strongly convex F) linear rate $\rho = \frac{1}{\kappa}$ (becomes with catalyst $\approx \frac{1}{\sqrt{\kappa}}$) for catalyst

(F convex case) $\quad O(\frac{1}{t})$ on F (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 ...



14h15

non-convex optimization

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

FW gap

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

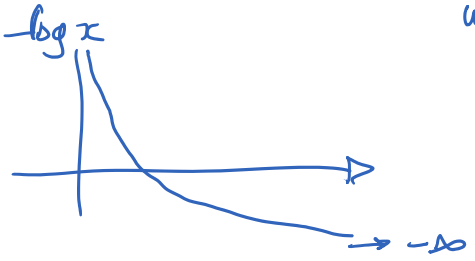
$(1) \to \frac{1}{\epsilon}$ Nesterov $\frac{1}{\sqrt{\epsilon}}$

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

btw: if f is μ-strongly convex

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

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



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

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

$$w_{t+1} = w_t - \frac{1}{L}\nabla f(w_t)$$

(L-smoothness of f but no need of convexity)

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

if f^* is finite

$$\Rightarrow f(w_T) - f^* \le f(w_0) - f^* - \frac{1}{2L}\sum_t \|\nabla f(w_t)\|^2$$

$$\Rightarrow \sum_t \|\nabla f(w_t)\|^2 \le 2L\,(f(w_0) - f^*)$$

$$\Rightarrow T \cdot \min_{s \le T} \|\nabla f(w_s)\|^2 \le 2L\,(f(w_0) - f^*)$$

ie. $$\boxed{\min_{s \le T} \|\nabla f(w_s)\|^2 \le \frac{2L}{T}(f(w_0) - f^*)}$$

NIPS 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)

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

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

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

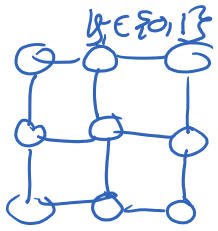
Submodular optimization

Submodularity is an analog of convexity 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 $E(y) = \sum_i \theta_i y_i - \sum_i \sum_{j \in \text{neighbor of } i} \theta_{ij} y_i y_j$

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

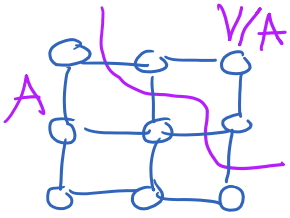
$F(A_y) =$ ↗ when $\theta_{ij} > 0$, $\Rightarrow E(y)$ is submodular

"attractive potential"

MRF here is "associative Markov network"

AMN

can minimize $E(y)$ (or $F(A_y)$) by using "graph cut algorithm"



equivalence with min cost network flow problem

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

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

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

"diminishing return property"

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

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

* link with convexity $\rightarrow$ Lovasz extension (def, set) $A \subseteq V$

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

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 $w = \sum_i \alpha_i v_i \Rightarrow f(w) = \sum_i \alpha_i F(A_i)$ ($v(A_i)$)

F is submodular $\Leftrightarrow$ Lovasz extension f is convex

* Can write $f(w) = \max_{s \in B(F)} \langle s, w \rangle$ ← this can be computed efficiently using greedy alg

"Base polytope"

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

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

$\partial f(w_t) = \operatorname{argmax}_{s \in B(F)} \langle s, w_t \rangle$

* with l_2-regularization, use duality to get a smooth obj

$\min_{s \in B(F)} \frac{1}{2} \|s\|^2$

→ use "min-norm pt." alg variant of FCFW alg. * SOTA for submodular opt.