Lecture 21 - Catalyst

Thursday, April 4, 2024 9:27 AM

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 approximate inner loop of prox-operator

proximal point algorithm: is proximal gradient with $f=0$

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

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

Catalyst alg. (for μ-strongly $F(w)$)

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

repeat:

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

ε_t: to be specified

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

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.: $\alpha_{t+1}^2 = (1-\alpha_{t+1})\alpha_t^2 + q\alpha_{t+1}$

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

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

e.g. $\alpha_0 = \sqrt{q}$

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:

↳ strong convexity $G_t(w)$

acceleration results :

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

then with correct constants for γ & schedule for ε_t

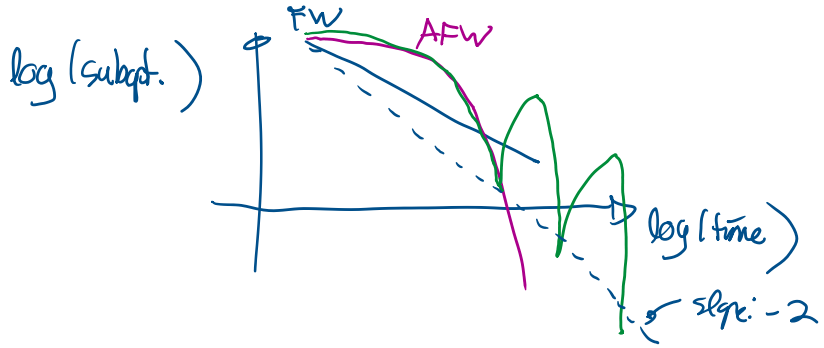
(μ-strongly convex F) before: linear rate $\rho = 1 - \frac{1}{\kappa}$ becomes with catalyst $\approx 1 - \frac{1}{\sqrt{\kappa}}$ for catalyst

(F convex case) $\qquad O\left(\frac{1}{\varepsilon}\right)$ on F becomes $O\left(\frac{1}{\varepsilon^2}\right)$

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

Issue: catalyst is not adaptive to local strong convexity

and finicky for choice of γ, ε_t, μ etc...



10h15
10h35

non-convex optimization

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

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

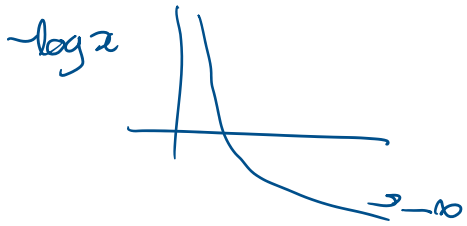
GD $\to \frac{1}{\varepsilon}$ $\qquad$ Nesterov $\frac{1}{\sqrt{\varepsilon}}$

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

a) btw: if f is μ-strongly convex

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

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



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

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

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

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

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

if f^* is finite

$$\Rightarrow f(w_{t+1}) - f^* \le f(w_0) - f^* - \frac{1}{2L}\sum_{s=0}^{t}\|\nabla f(w_s)\|^2$$

$$\Rightarrow \sum_{s=0}^{t}\|\nabla f(w_s)\|^2 \le 2L\,(f(w_0) - f^*)$$

$$(t+1)\cdot \min_{s \le t}\|\nabla f(w_s)\|^2 \le 2L\,(f(w_0) - f^*)$$

i.e. $$\boxed{\min_{0 \le s \le t}\|\nabla f(w_s)\|^2 \le \frac{2L}{t+1}\,(f(w_0) - f^*)}$$ //

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	$O\left(\min\left(\frac{1}{\epsilon^2}, \frac{n^{2/3}}{\epsilon}\right)\right)$

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

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	$O\left(\min(\frac{1}{\epsilon^2}), \frac{n^{2/3}}{\epsilon}\right)$	—

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.)

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

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

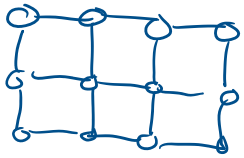
$2^V = \{V \to \{0,1\}\}$ = set of all subsets of V

Concrete example:

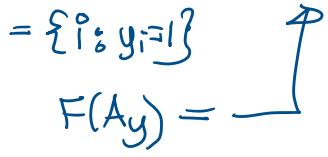
Ising model

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

$y_i \in \{0,1\}$



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

$F(A_y) =$ 

when $\theta_{ij} > 0$ ⇒ $E(y)$ is submodular

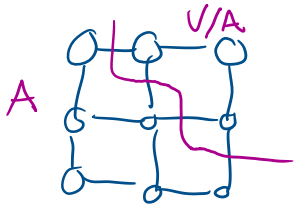
"attractive potential"

MRF here is "associative MRF"

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

equivalence with min cost network-flow

(or $F(A_y)$)



equivalence with min cost 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

"diminishing return property"

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

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

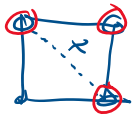
($|A|$: cardinality)

* link with convexity $\rightarrow$ Lovasz extension (cts. fct.)

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

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

(because convex fct. on $[0,1]^d$)



$f(w) = F(A_w)$ when $w = v(A)$

let's say $w = \sum_i \alpha_i v_i$ ($v_i = v(A_i)$) $\Rightarrow f(w) = \sum_i \alpha_i F(A_i)$

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

(Turns out)

* can write $f(w) = \max_{s\in B(F)} \langle s, w\rangle$ ← "Base polytope"

* this can be computed efficiently using greedy alg.

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

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

→ 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.

→ use "min-norm pt." alg, variant of FCFW alg.

⊗ SOTA for submodular min.