Lecture 21 - catalyst

Tuesday, March 31, 2020 14:27

today : · finish prox example
· catalyst → accelerate non-convex opt.
· submodular opt.

finish prox example

* to be useful, need $\text{prox}_\gamma^\Omega$ to be efficiently computable

$$\text{prox}_\gamma^{\|\cdot\|_1}(z) = \underset{w}{\text{argmin}} \ \|w\|_1 + \frac{1}{2\gamma}\|w-z\|^2$$

"soft-thresholding" (component-wise) $\triangleq \begin{cases} \text{sgn}(z_d)[|z_d|-\gamma] & \text{if } |z_d| \geq \gamma \\ 0 & \text{o.w.} \end{cases}$

used esp. for lasso : ℓ_1-reg. least-square

FISTA → accelerated prox. gradient method

↳ SOTA for batch lasso

* scikit-learn → uses SAGA for lasso & ℓ_1-reg log. reg.

$$\min_w \frac{1}{n}\sum_i f_i(w) + \Omega(w)$$

prox SAGA : $w_{t+1} = \text{Prox}_\gamma^\Omega\left(w_t - \gamma\left[\nabla f_{i_t}(w_t) - g_{i_t}^t + \frac{1}{n}\sum_{j=1}^n g_j^t\right]\right)$

Could accelerate using "Catalyst"

Catalyst algorithm [Lin, Mairal & Harchaoui NIPS 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 pt. alg. : is proximal gradient with $f=0$

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

Catalyst alg. (for μ-strongly convex $F(w)$)

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

repeat:

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

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

using inner loop optimization alg. [e.g. SAGA or AFW]

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

[accelerated Nesterov trick piece]

"extrapolation"

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

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

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

Catalyst trick: use γ & ε_t s.t. overall # of inner loop calls give overall acceleration

with clever analysis of warm starting

acceleration results:

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

then with correct constants

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

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

results: can get (theory) accelerated SAGA

" SVRG

" AFW

etc...

15h27

AFW

non-convex optimization

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

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

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

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

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

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

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)

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

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

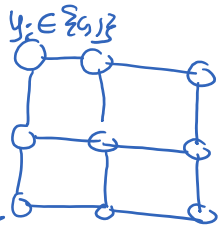
for set functions (combinatorial opt.)

$F: 2^V \to \mathbb{R}$ convention here: $F(\emptyset) = 0$

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

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

concrete example:



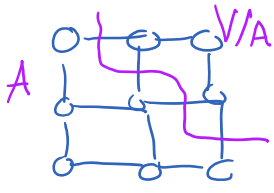
Ising model $y_i \in \{0,1\}$ $E(y) = \sum_i \theta_i y_i - \sum_i \sum_{j \text{ neighbor of } i} \theta_{ij} y_i y_j$

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

MRF here is called "associative Markov network" AMN

$F(A_y)$ where $A_y = \{i : y_i = 1\}$

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



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

$\Leftrightarrow$ 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 $B\subseteq A$

"diminishing return property"

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

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

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

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

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

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

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

  F is submodular $\Leftrightarrow$ Lovasz extension is convex

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

  $B(F)$: "Base polytope"

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

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

* with ℓ_2-regularization, use duality to get a smooth obj. $\min_{s\in B(F)} \frac{1}{2}\|s\|^2$

  $\rightarrow$ use "min-norm pt." alg. variant FCFW alg. $\ominus$ SOTA for submodular opt.