Lecture 21 - catalyst

Tuesday, April 2, 2019 14:35

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) $= \begin{cases} \text{sgn}(z_i)\,[|z_i| - \gamma] & \text{if } |z_i| \geq \gamma \\ 0 & \text{o.w.} \end{cases}$

used e.g. for lasso : ℓ_1-reg. least-square

FISTA → accelerated prox. gradient method

↳ state-of-the-art for batch lasso

* scikit-learn → use SAGA for lasso [next class]

could accelerate using "catalyst"

Catalyst algorithm [Lin, Mairal & Harchaoui NIPS 2015]

"meta-algorithm": outer loop which uses a linearly convergent alg. in the 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 $f=0$

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

catalyst alg. [for μ-strongly convex $F(w)$]

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

repeat:

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

(ε_t: to be specified)

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

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]

"extrapolation"

like a "momentum"

"extrapolation"

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

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

  (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 $\gamma \leq \epsilon_t$

s.t. overall # of inner loop calls give an overall acceleration

with clever analysis of warm starting

acceleration results:

if inner loop alg. has convergence expl $(-\frac{\tilde{\mu}}{L}t)$ ($\tilde{\mu}$: strong convexity for Gt.) $\tilde{\mu} \geq \mu + \frac{1}{\gamma}$

then with correct constants:

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

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

result & can get accelerated SAGA
" SVRG
" AFW
etc...

15h22

Non-convex optimization

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

$g(w_t)$: FW-gap

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

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

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	$O\left(\frac{1}{\epsilon^2}\right)$
GD	$O\left(\frac{n}{\epsilon}\right)$
SVRG	$O\left(n+\frac{n^{2/3}}{\epsilon}\right)$
SAGA	$O\left(n+\frac{n^{2/3}}{\epsilon}\right)$
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

Submodular optimization

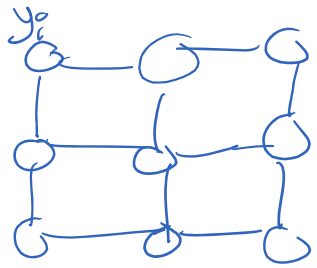
Submodularity is an analog of convexity for tractability for set functions (combinatorial opt.)

convention here: $F(\emptyset) = 0$

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

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

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



Concrete example: Ising model $y_i \in \{0,1\}$

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

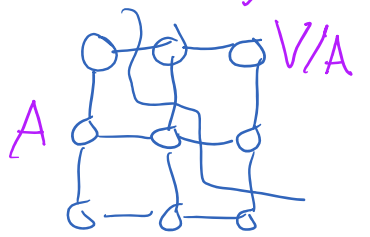
when $G_{ij} \geq 0$, $E(y)$ is submodular

"attractive potential"

MRF here is called "associative Markov network" AMN

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

can minimize this using "graph cut"



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

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

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

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

$B \subseteq A$

"diminishing return property"

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

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

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

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

Lovasz extension f extends $F(A)$ & can caress the whole 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 f is convex

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

"Base polytope"

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

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

→ use "min-norm pt" alg.

variant FCFW alg.

(*) S.O.T.A. for submodular minimization