Lecture 21 - scribbles - more optimization

Friday, March 31, 2017
15:02

- SAG for CRF
- proximal method, acceleration & non-convexity

SAG for CRF:

$$-\nabla f_i(w) = \sum_c \sum_{\tilde{y}_c} p(\tilde{y}_c | x^{(i)}, w)\, \psi_i(\tilde{y}_c)$$

$\Rightarrow$ store $\left(p(\tilde{y}_c|x^{(i)}, w)\right)_{\tilde{y}_c, c}$ in memory for SAG/SAGA correction

uses other tricks: • NUS
• line search on step-size $\Rightarrow$ this requires marginal inference "oracle call" for every step-size tried

SAG for CRF $\quad w^{(t+1)} = (1-\lambda\gamma_t)\, w^{(t)} - \gamma_t \left[\underbrace{\nabla f_i(w^{(t)}) - g_i^t}_{n} + \frac{1}{n}\sum_j g_j^t\right]$

OEG $\quad \alpha_i(\tilde{y})^{(t+1)} \propto \alpha_i(\tilde{y})^{(t)} \exp\left(-\gamma_t \nabla_{\alpha_i(\tilde{y})} D(\alpha^{(t)})\right)$

SDCA (stochastic dual coordinate ascent) $\quad$ do $\alpha_i(\tilde{y})^{(t+1)} = (1-\gamma_t)\alpha_i(\tilde{y})^{(t)} + \gamma_t \underbrace{\tilde{S}_i(\tilde{y})^{(t)}}_{\text{get this}}$ using "marginal inference"

$$w^{(t)} \overset{\text{KKT relationship}}{=} A\alpha^{(t)}$$

$$\left[\cdots \frac{1}{\lambda n}\psi_i(\tilde{y}) \quad\right]$$

[roughly: line search on $H(\alpha_\gamma)$ $\leftarrow$ 1d line search]

Proximal gradient method:

· generalization of projected gradient method to other non-smooth functions

→ generalization of projected gradient method to other non-smooth functions

<u>composite framework</u> : $F(w) = f(w) + \Omega(w)$ where f is convex and L-smooth

but Ω is convex, not necessarily smooth

constrained opt. setting : $\Omega(w) = \underbrace{\delta_M(w)}^{\text{indicator of } M} \triangleq \begin{cases} 0 & \text{if } w \in M \\ +\infty & \text{o.w.} \end{cases}$

ℓ_1 - regularization : $\Omega(w) = \lambda \|w\|_1$

<u>proximal gradient algorithm</u> :

$$\boxed{w_{t+1} = \arg\min_w \underbrace{f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + \frac{1}{2\gamma_t} \|w - w_t\|^2}_{B_t(w)} + \Omega(w)}$$

if $\gamma_t \leq \frac{1}{L}$; then $f(w) \leq B_t(w) \quad \forall w$

we can rewrite $B_t(w) = \frac{1}{2\gamma_t} \|w - [w_t - \gamma_t \nabla f(w_t)]\|^2 + cst.$ (by completing the square)

aside:

optimality condition on $F(w)$:

$w^* = \text{prox}_\gamma^\Omega(w^* - \gamma \nabla f(w^*))$ $\forall \gamma > 0$

$$\boxed{w_{t+1} = \text{Prox}_{\gamma_t}^\Omega (w_t - \gamma_t \nabla f(w_t))}$$

inf convolution

$(\Omega \,\square\, \|\cdot\|^2)(z) = \inf_w \Omega(w) + \|w - z\|^2$

"proximal operator" : $\text{prox}_\gamma^\Omega(z) \triangleq \arg\min_w \left\{ \Omega(w) + \frac{1}{2\gamma} \|w - z\|^2 \right\}$

strongly convex ⇒ unique minimizer

like for project, prox is <u>non-expansive</u> i.e. 1-Lipschitz

i.e. $\|\text{prox}_\gamma^\Omega(w) - \text{prox}_\gamma^\Omega(w')\|_2 \leq \|w - w'\|_2$

⇒ rates for unconstrained gradient method transfer to proximal analog

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

$$\mathrm{prox}_\gamma^{\|\cdot\|_1}(z) = \arg\min_w \|w\|_1 + \frac{1}{2\gamma}\|w-z\|^2$$

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

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 in inner loop of prox operator

proximal point algorithm: is proximal gradient with $f \equiv 0$

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

Catalyst alg.: (for μ-strongly convex F)

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

repeat:

any w_{t+1} — to be specified

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

$\triangleq G_t(w)$

(using our inner alg. [e.g. SAGA, FW ...])

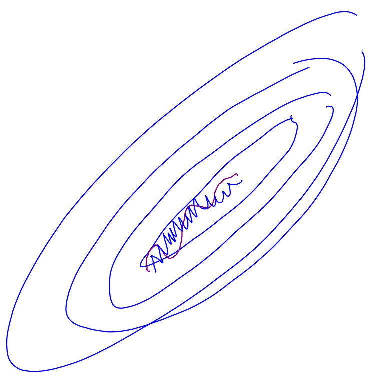
(accelerated Nesterov trick piece): $z_{t+1} = w_{t+1} + \beta_{t+1}\underbrace{(w_{t+1} - w_t)}_{\text{like a 'momentum'}}$

extrapolation

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

solve for α_{t+1} in equation $\alpha_{t+1}^2 = (1-\alpha_{t+1})\alpha_t^2 + q\alpha_{t+1}$
(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 is: use γ and ε_t

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

with clever analysis of warm starting

result is if inner loop alg. has convergence $\exp(-\frac{\mu}{L} t)$ strong convexity of inner loop problem

then with correct constants

go from $\frac{1}{t} \to \frac{1}{t^2}$ for outer loop convex minimization

$\frac{1}{k} \to \approx \frac{1}{\sqrt{k}}$ for strongly-convex case

Non-convex optimization:

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

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

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

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

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

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

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

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

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

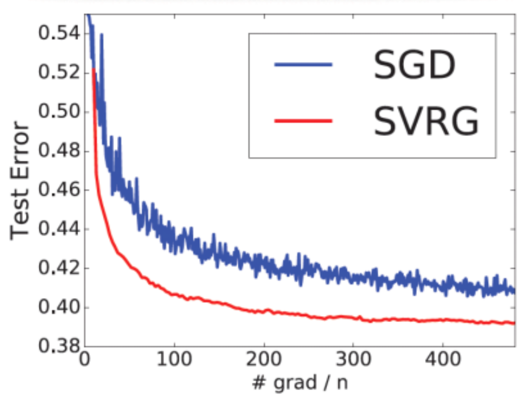
$$g(\theta) - g(\theta^*) \leq \frac{1}{2\mu}\|\nabla g(\theta)\|^2 \qquad \mathbb{E}[g(\theta_t) - g^*] \leq \epsilon$$

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

Empirical results



CIFAR10 dataset; 2-layer NN

Pointers

- SAG for CRF paper:
  - Non-Uniform Stochastic Average Gradient Method for Training Conditional Random Fields
    M. Schmidt, R. Babanezhad, M.O. Ahmed, A. Defazio, A. Clifton, A. Sarkar
    AISTATS 2015 | code
- Proximal gradient method
  - see slides of this great optimization class by L. Vandenberghe
- Catalyst -- meta-algorithm for acceleration:
  - A Universal Catalyst for First-Order Optimization
    Hongzhou Lin, Julien Mairal, Zaid Harchaoui
    NIPS 2015
- non-convex optimization:
  - see slides from Suvrit Sra at the NIPS 2016 tutorial on "Large-Scale Optimization: Beyond Stochastic Gradient Descent and Convexity" (e.g. table of rates on p. 20 and 22)

Other good optimization pointers:

- great coverage of convex optimization by Mark Schmidt at the Machine Learning Summer School in 2015) - slides | video
- two great classes on optimization:
  - EE236C - Optimization Methods for Large-Scale Systems (Spring 2016) - Prof. L. Vandenberghe, UCLA - link
  - Convex optimization class by Rayn Tibshirani at CMU - Fall 2015 - link
    - slides on the Frank-Wolfe lecture
    - nice summary table
- tutorial by Francis Bach on SAG et al. at NIPS 2016 -- slides