Lecture 18 - scribbles - variance reduced SGD

Friday, March 23, 2018 14:16

today's: · variance reduced SGD
· CRF

variance reduced SGD

setup: $\min_x \underbrace{\frac{1}{n}\sum_i f_i(x)}_{\triangleq f(x)}$ where f is μ-strongly convex, L-smooth

batch gradient method: $x_{t+1} = x_t - \gamma \frac{1}{n}\sum_i \nabla f_i(x_t)$ (∇f) — $O(n)$ to compute

$\gamma = \frac{1}{L}$ $\quad f(x_t) - f^* \leq (1-\rho)^t (f(x_0) - f^*)$ — linear rate

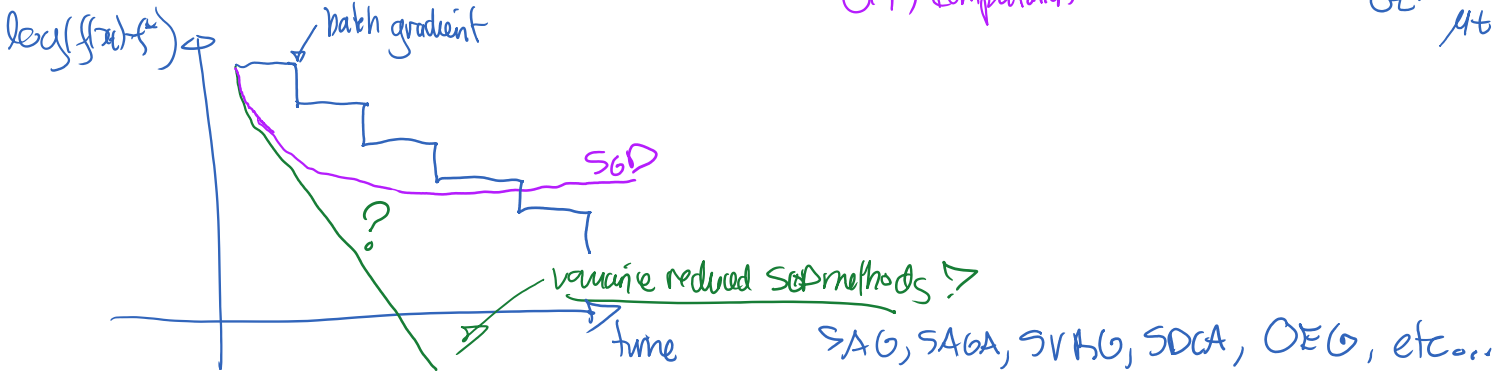
where $\rho \approx \frac{\mu}{L} = \frac{1}{\kappa}$ $\quad \kappa \triangleq \frac{L}{\mu}$ condition #

stochastic gradient method [Robbins & Monro 1951]
incremental gradient method

$x_{t+1} = x_t - \gamma_t \nabla f_{i_t}(x_t)$

where $i_t \sim \text{unif}\{1,\dots,n\}$ — $O(1)$ computation

- $\gamma_t = \text{const } \gamma$ → linear rate up to a ball of radius γ, rate $\rho \propto \gamma$
- $\gamma_t \sim \frac{1}{\mu t}$ → $O(\frac{1}{\mu t})$ rate (sublinear)



SAG, SAGA, SVRG, SDCA, OEG, etc...

SAG (stochastic average gradient) [LeRoux, Schmidt & Bach 2012]

SAG: • store past gradients for each i
• update one at step t

SAG $\begin{cases} \text{pick } i_t \sim \text{unif}; \text{ update } g_{i_t}^{(t+1)} \triangleq \nabla f_{i_t}(x_t); \; g_j^{(t+1)} = g_j^{(t)} \; \forall j \neq i_t \\ x_{t+1} = x_t - \gamma \frac{1}{n} \sum_{i=1}^{n} g_i^{(t+1)} \end{cases}$

stale gradients

is an approximation of $\nabla f(x_t)$

$O(1)$ cost per iteration
(but $O(n)$ storage cost)

big surprise: converge linearly and fast

"increment aggregated gradient" (IAG) [Blatt & al. 2007] where you cycle deterministically through $\{1, \dots, n\}$

→ linear rate for quadratic function but $\gamma_{max} \simeq O(\frac{1}{nL})$

"big stepsize"

SAG convergence rate: thm: with $\gamma_t = \frac{1}{16L}$ where $L = \max_i$ (Lipschitz (∇f_i))

$$\mathbb{E} f(x_t) - f^* \leq \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8n}\right\}\right)^t C_0$$

constant

i.e.

$$\rho_{SAG} = \min\left\{\frac{1}{16\kappa_{SAG}}, \frac{1}{8n}\right\}$$ compare with $\rho_{Grad} \approx \frac{1}{\kappa_{grad}}$

example: log regression on RCV1

$n = 700k \quad L = 0.25 \quad \mu = \frac{1}{n} \quad (\Rightarrow \kappa = \frac{n}{4})$

rate comparison:

gradient method $\left(\frac{L-\mu}{L+\mu}\right)^2 = 0.99998$

accelerated Grad. (Nesterov) $\left(1 - \sqrt{\frac{\mu}{L}}\right) = 0.99761$

Nesterov lower bound: has rate $\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2 = 0.99048$

SAG: (n iterations) $(1-\rho_{SAG})^n = 0.88250$

practical aspects (see Schmidt journal paper)

input data $(x_i, y_i)_{i=1}^n$

a) storage is $f_i(w) = h(x_i^T w) \Rightarrow \nabla f_i(w) = \underbrace{h'(x_i^T w)}_{\text{scalar}} x_i$

instead of $O(nd)$ storage $\rightsquigarrow O(n)$

b) initialization? Just is run SGD for one pass, then start SAG/SAGA etc...
of g_i's memory

c) step-size?
- $1/L$
- cheap line search heuristic (comes from FISTA)

$\{$ while $f_i\left(w_t - \frac{1}{\tilde{L}_i}\nabla f_i(w_t)\right) \geq f_i(w_t) - \frac{1}{2\tilde{L}_i}\|\nabla f_i(w_t)\|^2$
set $\tilde{L}_{i,new} = 2\tilde{L}_{i,old}$

$\tilde{L}_i$ is surrogate for L_i

use step-size $\frac{1}{\tilde{L}_i}$

d) non-uniform sampling?

sample $i \sim \frac{L_i}{\sum_j L_j}$

use step-size $\frac{1}{L_i}$

e) stopping criterion?

you can use $\frac{1}{n}\sum_j g_j^t$ as approximate $\nabla f(w_t)$

f) sparse features?

$$w_{t+1} = w_t - \gamma\left(\nabla f_{i_t}(w_t) - g_i^t + P_{S_i}\left(\frac{1}{n}\sum_j g_j^t\right)\right)$$ (sparse SAGA)

($\nabla f_{i_t}(w_t)$: sparse; $\frac{1}{n}\sum_j g_j^t$: dense)

[Leblond et al. 2017]

P_{S_i}: weighted projection on support of x_i

$S_i \triangleq \{u : (x_i)_u \neq 0\}$

15h47

Variance reduction idea: X & Y are r.v.

goal: estimate $\mathbb{E}X$ using M.C. samples

suppose: $\mathbb{E}Y$ is cheap to compute and Y is correlated with X

consider estimator $\Theta_\alpha \triangleq \alpha(X-Y) + \mathbb{E}Y$ to approximate $\mathbb{E}X$

$\alpha \in [0,1]$

properties: $\mathbb{E}\Theta_\alpha = \alpha\mathbb{E}X + (1-\alpha)\mathbb{E}Y$ ⇝ unbiased (i.e. $\mathbb{E}\Theta_\alpha = \mathbb{E}X$)

if $\mathbb{E}Y = \mathbb{E}X$ [not interesting]

or $\alpha = 1$

variance $\mathrm{Var}(\Theta_\alpha) = \alpha^2[\mathrm{var}(X) + \mathrm{var}(Y) - 2\,\mathrm{cov}(X,Y)]$

variance $\mathrm{Var}[\alpha X + (\mathbb{E}Y - Y)] = \alpha^2[\mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y)]$

variance reduction

for $\alpha = 1$ (unbiased) $\quad \Theta_\alpha = X + (\mathbb{E}Y - Y)$

correction

SGD setting: X is $\nabla f_i(x_t)$ $\quad \mathbb{E}X$ = batch gradient

SAG/SAGA algorithm: Y is g_i [past stored gradient]

$$\mathbb{E}Y = \frac{1}{n}\sum_i g_i$$

SAG algorithm: $\alpha = \frac{1}{n}$ (biased)

SAGA " : $\alpha = 1$ (unbiased)

SAG: $$w_{t+1} = w_t - \gamma \Big[\underbrace{\nabla f_{i_t}(w_t)}_{X} - \underbrace{g_{i_t}^t}_{Y} + \underbrace{\frac{1}{n}\sum_j g_j^t}_{\mathbb{E}Y} \Big] \quad \text{(biased)}$$

$\alpha = \frac{1}{n}$

SAGA: $$w_{t+1} = w_t - \gamma \Big[\nabla f_{i_t}(w_t) - g_{i_t}^t + \frac{1}{n}\sum_j g_j^t \Big] \quad \text{(unbiased)}$$

$\alpha = 1$?

SVRG: (stochastic variance reduced gradient) $$w_{t+1} = w_t - \gamma \Big[\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{old}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w_{old}) \Big] \quad \text{(unbiased)}$$

w_{old} is updated from outer loop

CRF objective

$\max_{\tilde{y}} \ell_i(\tilde{y}) - w^T \psi_i(\tilde{y})$

primal	dual
SVM struct: $\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_i H_i(w)$ (hinge)	$\max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\frac{\lambda}{2}\|w(\alpha)\|^2 + \frac{1}{n}\sum_i \ell_i^T \alpha_i$
CRF: $\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_i -\log p(y^{(i)}\|x^{(i)};w)$	$\max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\frac{\lambda}{2}\|w(\alpha)\|^2 + \frac{1}{n}\sum_i H_i(\alpha_i)$ (entropy)
$\log(\sum_{\tilde{y}} \exp(-w^T \psi_i(\tilde{y})))$	$\triangleq -\sum_{\tilde{y}} \alpha_i(\tilde{y}) \log \alpha_i(\tilde{y})$

KKT $\rightarrow$ $w(\alpha) = \frac{1}{\lambda n}\sum_i \sum_{\tilde{y}} \alpha_i(\tilde{y}) \psi_i(\tilde{y})$

$= \frac{1}{\lambda n}\sum_i \sum_c \mu_{i;c}(\tilde{y}_c) \psi_{i,c}(\tilde{y}_c)$ (c: from MRF)

$p(y|x;w) \propto \exp(\langle w, \phi(x,y)\rangle)$

$\Rightarrow$ at optimality

$\alpha_i^*(y) = p(y \mid x^{(i)}; w^*)$

$\Rightarrow \alpha_i^* \in$ interior of $\Delta_{|\mathcal{Y}_i|}$

unlike sparse solution in structured SVM

CRF optimization:

SAG for CRF

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

OEG
online exponentiated gradient

$$\alpha_{i_t}(\tilde{y})^{(t+1)} \propto \alpha_{i_t}(\tilde{y})^{(t)} \exp(-\gamma_t \nabla_{\alpha_{i_t}(\tilde{y})} D(\alpha^{(t)}))$$

neg. dual objective

EG alg $\rightarrow$ proximal gradient step using $KL(\alpha\|\alpha_t)$ as Bregman divergence

prox term

SDCA
Stochastic dual coordinate ascent

$$\alpha_{i_t}(\tilde{y})^{(t+1)} = (1-\gamma_t)\alpha_{i_t}(\tilde{y})^{(t)} + \gamma_t \hat{s}_{i_t}(\tilde{y})^{(t)}$$

get this using marginal inference

$p(\tilde{y} | x_{i_t}; w(\alpha_t))$

[note: BCFW is special case of SDCA on SVMstruct]