Lecture 20 - variance reduction

Thursday, March 26, 2020 13:24

today: · variance reduction perspective
· application to CRF

variance reduction idea

X & Y are R.V.'s

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 $\alpha \in [0,1]$ $\quad \theta_\alpha \triangleq \alpha(X-Y) + \mathbb{E}Y$ to approximate $\mathbb{E}X$

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 $\alpha = 1$ $\mathbb{E}Y \neq \mathbb{E}X$ [not interesting]

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

→ variance reduction?

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

SGD setting:

X is $\nabla f_i(x_t)$; $\mathbb{E}X$ = batch gradient

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

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

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

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

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

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

$$\text{SVRG: } x_{t+1} = x_t - \gamma\left[\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_{old}) + \frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{old})\right] \quad \text{(unbiased)}$$

(stochastic variance)

SVRG : $x_{t+1} = x_t - \gamma \left[\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_{old}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(x_{old}) \right]$ (unbiased)

(stochastic variance reduced gradient)

x_{old} is updated from outer loop

SVRG algorithm

for $k=0, \ldots$ (outer loop)

compute $g_{ref} \triangleq \frac{1}{n}\sum_j \nabla f_j(x^{(k)})$

for $t=0, \ldots, T_{max}$

sample i_t

$x_{t+1}^{(k)} = x_t^{(k)} - \gamma \left[\nabla f_{i_t}(x_t^{(k)}) - \nabla f_{i_t}(x^{(k)}) + g_{ref} \right]$

end

$x^{(k+1)} = x_{T_{max}}^{(k)}$

end

questions:

- what is T_{max}
- what is γ?

original SVRG convergence result: need $\gamma \leq \approx \frac{1}{L}$

$T_{max} \geq \approx \frac{L}{\mu} = K$ → to run alg., need to know K;

⇒ not adaptive to local strong convexity

tweak of SVRG (now called "loopless") [Hoffmann & al. NIPS 2015] $T_{max} \sim Geom(\_)$

[at inner loop iteration, do a batch gradient comp. with prob $\frac{1}{n}$]

then, get same convergence result as SAGA

↳ size of inner loop $\mathbb{E}[T_{max}] = n$; overall cost of SVR $\approx 3 \cdot$ (SGD for n updates)

SAG/SAGA, loopless SVRG, convergence for convex fct. ($\mu = 0$) get $\min_t \{\mathbb{E} f(x_t) - f^*\} = O\left(\frac{1}{t}\right)$

[contrast with $1/\sqrt{t}$ for SGD]

15h34

CRF optimization

$\max_{\tilde{y}} \left(l_i(\tilde{y}) - w^T \psi_i(\tilde{y}) \right)$

CRF optimization

CRF objective:

	primal	dual
SVMstruct	$\min_w \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^n H_i^{\text{hinge}}(w)$	$\max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\frac{\lambda\|w(\alpha)\|^2}{2} + \frac{1}{n}\sum_{i=1}^n \ell_i^T \alpha_i$
CRF	$\min_w \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^n -\log p(y^{(i)}\|x^{(i)}; w)$	$\max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\frac{\lambda\|w(\alpha)\|^2}{2} + \frac{1}{n}\sum_{i=1}^n H_i^{\text{entropy}}(\alpha_i)$

where $H_i^{\text{hinge}}(w) = \max_{\tilde{y}} (\ell_i(\tilde{y}) - w^T \psi_i(\tilde{y}))$

$\log\left(\sum_{\tilde{y}} \exp(-w^T \psi_i(\tilde{y}))\right)$

$H_i^{\text{entropy}}(\alpha_i) \triangleq \sum_{\tilde{y}} \alpha_i(\tilde{y}) \log \alpha_i(\tilde{y})$

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

$= \frac{1}{\lambda n}\sum_i \sum_{c \in \mathcal{C}} \mu_{i,c}(\tilde{y}_c) \psi_{i,c}(\tilde{y}_c)$ (from MRF)

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

(*) at optimality: $\alpha_i^*(y) = p(y|x^{(i)}; w(\alpha^*))$

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

unlike sparse solution in structured SVM

CRF optimization:

- primal is smooth & strongly convex [vs. non-smooth for SVMstruct]
  - for a while, batch L-BFGS was method of choice [batch $\Rightarrow$ slow for large n]
  - [Collins & al. JMLR 2008]: online exponentiated gradient (OEG)

block-coordinate method on dual; exponentiated gradient step on block

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

(D: dual obj.)

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

$\hookrightarrow$ get linear conv. rate with cheap $O(1)$ updates (like SGD) [vs. $O(n)$ for batch method]

[can think of it as variance reduced method as well?]

- SAGA for CRF [Schmidt & al. AISTATS 2015]

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

$\in [0,1]$ (stochastic)

$g_i^{(t)}$ $\quad$ $\tilde{\alpha}_i^{(t)}$

2015]

- SDCA
  stochastic dual
  coord. ascent

$\alpha_{i_t}^{(t)}(\tilde{y}) = (1-\gamma_t)\alpha_{i_t}^{(t)}(\tilde{y}) + \gamma_t \tilde{S}_i^{(t)}(\tilde{y})$ $\quad \epsilon[0,1]$ (stabilize)

$\tilde{S}_i^{(t)}(\tilde{y}) \sim p(\tilde{y} | x_i; w(\alpha^{(t)}))$

$\triangleq \alpha(w)$

as a related fixed pt. update $\alpha_i^* = p(y|x_i; w(\alpha^*)) \; \forall i$

related to subgradient in primal

[note: BCFW is a special case of SDCA in S/M shut off obj. ?]

→ SOTA for CRF [LePriol & al. UAI 2018] (thanks to us)

proximal gradient method

↳ generalization of projected gradient method to other non-smooth fcts.

composite framework: $F(w) \triangleq f(w) + \Omega(w)$ where f is convex & L-smooth, Ω is convex but not nec. smooth

- constrained opt. $\Omega(w) = \delta_{\mathcal{M}}(w) \triangleq \begin{cases} 0 & \text{if } w \in \mathcal{M} \\ +\infty & \text{o.w.} \end{cases}$

"indicator on $\mathcal{M}$"

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

proximal gradient update:

$$w_{t+1} = \underset{w}{\operatorname{argmin}} \; \underbrace{f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + \frac{1}{2\gamma_t}\|w - w_t\|^2}_{\triangleq B_t(w)} + \Omega(w)$$

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

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

$\Rightarrow$ if $\Omega(w) = \delta_{\mathcal{M}}(w)$; we get the projected gradient alg.

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

$$w_{t+1} = \text{prox}^{\Omega}_{\alpha_t}(w_t - \alpha_t \nabla f(w_t))$$

↳ "proximal operator"

$$\text{prox}^{\Omega}_{\gamma}(z) \triangleq \arg\min_{w} \left\{ \Omega(w) + \frac{1}{2\gamma}\|w - z\|_2^2 \right\}$$

could replace with Bregman divergence to get other gener. (e.g. OEG)

(*) like projection, prox operator is non-expansive (i.e. 1-Lipschitz)

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

$\Rightarrow$ convergence rate of prox. grad. method on $F = f + \Omega$ are <u>same</u> as unconst. gradient descent on f