Lecture 19 - CRF & SAG

Friday, March 24, 2017
14:54

today: • CRF
, variance reduced SGD

$\psi_i(\tilde{y}) \triangleq \varphi(x^{(i)}, y^{(i)}) - \varphi(x^{(i)}, \tilde{y})$

CRF: log-loss $\ell(x,y;w) = -\log p(y|x;w)$ $\quad p(y|x;w) \propto \exp(\underbrace{s(x,y;w)}_{\langle w, \varphi(x,y)\rangle})$

suppose a MRF $s(x,y;w) = \sum_{C \in \mathcal{C}} s_C(x, y_C; w) = \sum_C \langle w, \varphi_C(x, y_C)\rangle$

optimization objective: primal dual $\quad \ell_i = [\ell(y^{(i)}, \tilde{y})]_{\tilde{y} \in \mathcal{Y}_i}$

SVM struct: $\min_w \frac{\lambda \|w\|^2}{2} + \frac{1}{n}\sum_i H_i^{\text{hinge}}(w)$ $\quad$ ($\max_{\tilde{y}} \ell_i(\tilde{y}) - w^T\psi_i(\tilde{y})$) $\qquad \max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\frac{\lambda\|w(\alpha)\|^2}{2} + \frac{1}{n}\sum_i \ell_i^T \alpha_i$

CRF: $\min_w \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_i \underbrace{-\log p(y^{(i)}|x^{(i)}, w)}_{\log\left(\sum_{\tilde{y}} \exp(-w^T\psi_i(\tilde{y}))\right)}$ $\qquad \max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\frac{\lambda\|w(\alpha)\|^2}{2} + \frac{1}{n}\sum_i H_i^{\text{entropy}}(\alpha_i)$

$\triangleq -\sum_{\tilde{y}} \alpha_i(\tilde{y}) \log \alpha_i(\tilde{y})$

KKT: $\boxed{w(\alpha) = \frac{1}{\lambda n}\sum_i \sum_{\tilde{y}} \alpha_i(\tilde{y})\psi_i(\tilde{y}) = \frac{1}{\lambda n}\sum_i \sum_C \sum_{\tilde{y}_C} \mu_{i,C}(\tilde{y}_C)\psi_{i,C}(\tilde{y}_C)}$

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

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

Unlike sparse solution in structured SVM

log2

binary entropy

1

$\nabla_w \text{CRF-primal}(w) = \lambda w + \frac{1}{n}\sum_i \nabla_w \underbrace{(-\log p(y^{(i)}|x^{(i)}; w))}_{\log Z_i(w) - w^T\varphi_i(y^{(i)})}$

$\mathbb{E}_{\tilde{y}|x^{(i)};w}[\varphi_i(\tilde{y})] - \varphi_i(y^{(i)})$

$\mathbb{E}_{\tilde{y}|x^{(i)},w}[-\psi_i(\tilde{y})]$ ← to compute this;

$$\nabla_w = 0 \Rightarrow w^* = \frac{1}{\lambda n}\sum_i \underbrace{\mathbb{E}_{\tilde{y}|x^{(i)},w^*}[\psi_i(\tilde{y})]}_{\sum_{\tilde{y}} p(\tilde{y}|x^{(i)},w^*)\,\psi_i(\tilde{y})}$$

$$\psi_i(\tilde{y}) = \sum_c \psi_{i,c}(\tilde{y}_c)$$

$$\mathbb{E}[\psi_i(\tilde{y})] = \sum_c \underbrace{\mathbb{E}[\psi_{i,c}(\tilde{y}_c)]}$$

$$\sum_c \sum_{\tilde{y}_c} \underbrace{p(\tilde{y}_c|x^{(i)},w)}\,\psi_{i,c}(\tilde{y}_c)$$

need to do marginal inference

→ need marginalization oracle

[vs a maximization oracle for SSVM struct]

Optimization for CRF:

primal objective is smooth and strongly convex [vs non-smooth for SVM struct]

- for a while, batch L-BFGS was method of choice [batch ⇒ slow for large n]
- [Collins & al, JMLR 2008]: online exponentiated gradient:

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

dual function

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

↝ get linear convergence rate with cheap $O(1)$ [vs. $O(n)$] updates (like SGD)

low hanging fruit: SDCA → have same ? convergence properties but with cheap line-search and AFAIK, have not been tried

$\boxed{\text{Variance reduced SGD}}$

minimize $f(w) = \frac{1}{n}\sum_i f_i(w)$ where f is μ-strongly convex, L-smooth (i.e. ∇f is L-Lipschitz)

batch gradient method $w_{t+1} = w_t - \gamma \nabla f(w_t)$ $\quad\sim \leq \exp(-\rho t)$

[Cauchy 1847]

$$\frac{1}{n}\sum_i \nabla f_i(w_t) \qquad O(n)$$

$\gamma = \frac{1}{L}$; $f(w_t) - f^* \leq (1-\rho)^t (f(w_0) - f^*)$

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

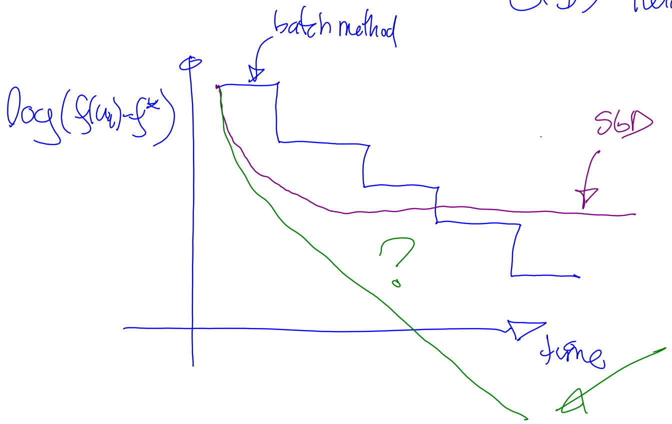
<u>stochastic gradient method</u> $\quad w_{t+1} = w_t - \gamma_t \nabla f_{i_t}(w_t)$, $\quad i_t \overset{\text{unif}}{\sim} \{1, \dots, n\}$

[Robbins & Monro 1951]

<u>incremental gradient method</u> $\rightarrow$ $O(1)$ iteration

$\gamma_t = \text{const } \gamma \rightarrow$ linear rate up to ball of radius γ ($\rho \propto \gamma$)

$\gamma_t \approx \frac{1}{\mu t} \rightarrow \tilde{O}\left(\frac{1}{\mu t}\right)$ rate <u>sublinear</u>



<u>variance reduced SGD methods</u>

<u>SAG (stochastic average gradient)</u> [Le Roux, Schmidt & Bach 2012]

standard batch gradient: $\quad w_{t+1} = w_t - \gamma \frac{1}{n}\sum_i \nabla f_i(w_t)$

store these "stale" gradients

SAG: pick i_t (unif at random); update $g_{i_t}^{(t+1)} \triangleq \nabla f_{i_t}(w_t)$; $g_j^{(t+1)} = g_j^{(t)}$ for $j \neq i_t$

and then $w_{t+1} = w_t - \gamma \sum_i g_i^{(t+1)}$

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

big surprise $\rightarrow$ this converges linearly and fast!

incremental aggregated gradient method (IAG) [Blatt et al 2007] where you go deterministically through i incremental aggregated gradient method (IAG) [Blatt et al. 2007] where you cycle deterministically through i_t

Convergence rate: thm.: with $\gamma_t = \frac{1}{16L}$ where $L = \max_i (\text{Lipschitz constant of } \nabla f_i)$

$$\boxed{\mathbb{E}[f(w_t)] - f^* \le \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8n}\right\}\right)^t C_0}$$

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: $n = 700k$, $L = 0.25$ $\mu = \frac{1}{n}$ $\left(\Rightarrow \kappa = \frac{n}{4}\right)$

Rate comparison

- Assume that $N = 700000$, $L = 0.25$, $\mu = 1/N$:
  - Gradient method has rate $\left(\frac{L-\mu}{L+\mu}\right)^2 = 0.99998$.
  - Accelerated gradient method has rate $\left(1 - \sqrt{\frac{\mu}{L}}\right) = 0.99761$.
  - SAG (N iterations) has rate $\left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8N}\right\}\right)^N = 0.88250$.
  - *Fastest possible* first-order method: $\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2 = 0.99048$.

Pointers

- Online exponentiated gradient for CRF paper:
  Collins, M., Globerson, A., Koo, T., Carreras, X., and Bartlett, P. L. Exponentiated gradient algorithms for conditional random fields and max-margin Markov networks.
  JMLR, 9:1775-1822, 2008
  (include a comparison with L-BFGS and also gives the dual of the CRF objective)
- stochastic average gradient (SAG):
  - original NIPS 2012 paper:
    N. Le Roux, M. Schmidt, F. Bach
    A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets.
    NIPS 2012 | code
  - massive journal version:
    M. Schmidt, N. Le Roux, F. Bach.

Minimizing Finite Sums with the Stochastic Average Gradient
Mathematical Programming, 162:83-162, 2017 | arxiv

- SAGA paper -- unbiased version of SAG (with simpler proof) as well as describe the related methods of SDCA and SVRG (will be covered next class) [see references therein for SDCA and SVRG...]
  - A. Defazio, F. Bach and S. Lacoste-Julien
    SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives
    NIPS 2014
  - for an even more bare bone proof, see:
    T. Hofmann, A. Lucchi, S. Lacoste-Julien, and Brian McWilliams
    Variance Reduced Stochastic Gradient Descent with Neighbors,
    NIPS 2015