Today: CRF, Variance Reduced SGD

CRF: \[ \log \text{-loss} \; s(a, y; w) = -\log p(y|x; w) \; \exp(s(x, y; w)) \]

Suppose a MTF \[ s(x, y; w) = \sum_{c \in S} s_c(x, y; w) = \sum_{c \in S} w_c \varphi(x, y) \]

Optimization objective:

SVM trick: \[ \begin{align*}
\min_w & \; \frac{1}{n} \sum_{i=1}^{n} s_i(x_i, y_i) \\
\text{subject to} & \; w_i^2 \leq H \quad (\forall i)
\end{align*} \]

CRF:

\[ \begin{align*}
\min_w & \; \frac{1}{2} \sum_{i=1}^{n} w_i^2 + \frac{1}{n} \sum_{i=1}^{n} s_i(x_i, y_i) \\
\text{subject to} & \; w_i^2 \leq H \quad (\forall i)
\end{align*} \]

KKT:

\[ \begin{align*}
\nabla w_i & = \frac{1}{n} \sum_{i=1}^{n} \nabla \log p(y|x; w) \\
\text{subject to} & \; w_i \leq \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i, y_i)
\end{align*} \]

\[ \text{Optimally:} \; \alpha^*_i(y) = p(y|x; w^*) \]

CRF-primal(w) = \[ \begin{align*}
\nabla w_i & = \frac{1}{n} \sum_{i=1}^{n} \nabla \log p(y|x; w) \\
\text{subject to} & \; w_i \leq \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i, y_i)
\end{align*} \]

Unlike sparse solution in structured SVM.
Optimization for CRFs:

- Objective is smooth and strongly convex [vs. non-smooth for SVMs]
- For a while, batch L-BFGS was method of choice [batch = slow for large n]
- [Collins et al., SMUR 2006] & online exponentiated gradient:
  - Block-coordinate method on dual & exponentiated gradient step on dual
  - $\text{Or}(y)^{(4n)} \propto \text{Or}(y)^{(4n)} \exp(-\gamma_b \nabla \cdot \mathbb{D}(\alpha^{(4n)}))$

  $\nabla \cdot \mathbb{D}(\alpha^{(4n)})$

  - Get linear convergence rate with cheap OR [vs. O(n)]

  - Low hanging fruit: SDCA + have some convergence properties
    - But with cheap line-search and AF AIK, have not been tried

Vanilla reduced SGD

- Minimize $f(\omega) = \frac{1}{n} \sum_i f_i(\omega)$ where $f$ is $\nu$-strongly convex, $L$-smooth (i.e., $\nabla^2 f$ is $L$-Lipschitz)

- Batch gradient method $\omega_{t+1} = \omega_t - \frac{1}{\gamma} \nabla f(\omega_t)$
Stochastic gradient method

\[ \omega_{t+1} = \omega_t - \beta \nabla_{\omega} f_t(\omega_t) \]

\[ p = \mu \leq 1 \quad \text{update} \]

\[ \mu_t \rightarrow 0(1) \text{ rate} \]

\[ \text{standard batch gradient} \quad \omega_{t+1} = \omega_t - \beta \nabla_{\omega} f_t(\omega_t) \]

\[ \text{SAO} \quad \sum_{i=1}^{n} \text{pick i update } g_{i}^{(t+1)} = \nabla f_{e_i}(\omega_{e_i}) \quad \omega_{t+1} = \omega_t - \beta \sum_{i=1}^{n} g_{i}^{(t+1)} \]

\[ \text{O(1) I/O when (but not) storage) \}

\[ \text{big surprise} \quad \text{this converges linearly and fast} \]

\[ \text{LeRoux, Schmidt \& Bach 2012} \]

\[ \text{SAO (stochastic average gradient)} \]
Pointers

- Online exponentiated gradient for CRF paper:
  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.

Rate comparison

- Assume that $N = 700000$, $L = 0.25$, $\mu = 1/N$:
  - Gradient method has rate $(\frac{1}{\sqrt{\mu}})^2 = 0.99998$. 
  - Accelerated gradient method has rate $(1 - \sqrt{\mu}) = 0.99761$. 
  - SAG ($N$ iterations) has rate $(1 - \min \{ \frac{1}{\mu L}, \frac{1}{2L} \})^N = 0.68250$. 
  - Fastest possible first-order method: $(\frac{\sqrt{\mu} - 1}{\sqrt{\mu} + 1}) = 0.99048$. 

Convergence rate: $f(n) = \frac{1}{10} + \sqrt{n}$, where $L = \max(\text{Lipschitz constant of } \nabla f)$. 

$E[f(x)] - f(x) \leq (1 - \min \{ \frac{\mu}{4L}, \frac{1}{8L} \})^n \cdot C_0$ 

i.e. $P_{\text{grad}} = \min \{ \frac{\mu}{4L}, \frac{1}{8L} \}$, compare with $P_{\text{grad}} = \frac{1}{K_{\text{grad}}}$. 

Example: $n = 70000$, $L = 0.25$, $\mu = \frac{1}{n}$, $\Rightarrow k = \frac{n}{4}$. 

\textit{Unbiased aggregated gradient method (IAG)} [Malek et al. 2007] where you cycle deterministically through it.
• 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