**Motivation and Overview of Contribution**

- Conditional random fields (CRFs) are a ubiquitous tool for structured prediction:
  - Allow the use of a high-dimensional feature set.
  - Formulated as convex optimization problem.
  - But very slow to train.

- Stochastic average gradient (SAG) methods are a new strategy for convex optimization:
  - Only look at a single training example on each iteration, like stochastic gradient methods.
  - Linear convergence rate similar to methods that process the entire dataset on every iteration.
  - Our contribution is applying SAG with non-uniform sampling (NUS) to CRFs.
  - We show how to reduce the memory requirements using structure in the gradient.
  - We propose a practical NUS scheme that substantially improves empirical performance.
  - We analyze the rate of convergence of the SAGA variant under non-uniform sampling.
  - SAG with NUS often outperforms existing methods for training CRFs.

**Conditional Random Fields (CRFs)**

- CRFs model probability of output $y \in \mathcal{Y}$ given input $x \in \mathcal{X}$ and features $F(x, y)$ using $p(y | x, w) = \exp \{ w^T F(x, y) \} / \sum_y \exp \{ w^T F(x, y) \}$.

- Given training examples $\{x_i, y_i\}$, standard approach is minimizing $\ell_2$-regularized NLL:
  $$
  \min_w \ell(w) = \frac{1}{n} \sum_i \log p(y_i | x_i, w) + \frac{\lambda}{2} \|w\|^2.
  $$

- Evaluating each $\log p(y_i | x_i, w)$ is expensive due to sum over $y_i$.

**Related Work on Deterministic, Stochastic, and Hybrid Methods**

- Deterministic gradient methods like L-BFGS [Wallach et al., 2002, Sha & Pereira, 2003]:
  - Require $O(n^2)$ iterations but $n$ gradient evaluations per iteration.
- Stochastic gradient methods [Vishwanathan et al., 2006, Finkel et al., 2008]:
  - Require $O(1/\epsilon)$ iterations but only a single gradient evaluation per iteration.
- Online exponentiated gradient [Collins et al., 2008]:
  - Requires $O(\log(1/\epsilon))$ iterations in terms of dual and dual gradient evaluation per iteration.
- Hybrid deterministic-stochastic [Friedlander & Schmidt, 2012]:
  - Requires $O(\log(1/\epsilon))$ iterations and growing number of gradient evaluations per iteration.

**Stochastic Average Gradient (SAG) for CRFs**

- Stochastic average gradient [LeRoux et al., 2012]:
  - Requires $O(\log(1/\epsilon))$ iterations and 1 gradient evaluation per iteration.

- SAG uses the iteration:
  $$
  w^{t+1} = \bar{w}^t - \frac{a_t}{\eta_t} \sum_i g_i.
  $$

- where $g_i = - \nabla \log p(y_i | x_i, w^t) + \lambda w^t$ for one randomly-chosen training example.

- Challenge is the memory required for storing the $g_i$:
  - $\nabla \log p(x_i | w)$ often sparse but depends on number of features.
  - $\lambda w$ is typically dense.

- Implementation issues for CRFs:
  - Sparse trick 1: to avoid storing $\lambda w$, use the exact gradient of the regularizer.
    $$
    w^{t+1} = (1 - \alpha)w^t - \frac{\alpha}{\eta} \sum_i g_i
    $$

  - Sparse trick 2: use representation $w^t = \sum_i x_i$ and "lazy updates" to avoid dense vector operations.
  - Stop size: we set $\alpha = 1/\lambda + \lambda$, and double approximation $\mathcal{L}_1$ when
    $$
    \ell(w - (1/\lambda)g) > \ell(w) + \frac{\alpha}{\eta} \|g\|^2.
    $$

**SAG with Practical Non-Uniform Sampling (NUS) Strategy**

- Assume each gradient has its own Lipschitz constant $L_i$, a value such that
  $$
  \|\nabla f_i(w) - \nabla f_i(v)\| \leq L_i \|w - v\|. \forall w, v,
  $$

- bounding how fast gradient $i$ can change.

- Key idea behind NUS: bias sampling probability $p_i$ towards Lipschitz constant $L_i$:
  - $p_i / \sum_j p_j$.
  - Convergence rate depends on $\max_i L_i$, instead of $L = \max_i L_i$.

- Practical partially-based strategy:
  - With probability $1/2$ choose $i$ uniformly.
  - With probability $1/2$ sample $i$ with probability $p_i / \sum_j p_j$.
  - Use a larger step-size of $\alpha' = 1/2 L_i / \sum_p$.
  - Initialize with $L_i = \mathcal{L}_1$ the first time an example is chosen.
  - Each time $i$ is chosen, set $L_i = 0.9 L_i$, then double it while $f(1)$ holds.

- If $f(1)$ holds $\ell$ times (without backtracking), do not change $L_i$ for 2$^{\ell+1}$ times next $i$ is sampled.

**Convergence Analysis for SAGA with Non-Uniform Sampling**

- We analyze a NUS extension of SAGA, which has similar performance but easier analysis.

- Let the sequences $\{w^i\}$ and $\{s^i\}$ be defined by
  $$
  w^i = w^i - \frac{a_i}{\eta_i} \sum_{j = 1}^i \nabla f_i(w^i) - \frac{1}{n} \sum_{j = 1}^i s^j
  $$

  $$
  s^i = \begin{cases}
  \nabla f_i(w^i) & \text{if } i = r_i, \\
  s^r_i & \text{otherwise}.
  \end{cases}
  $$

  where $j_i$ is chosen with probability $p_i$.

  - (a) If $i$ is chosen with probability $p_i = \frac{1}{i}$, then with $\alpha = \frac{n \eta_i}{\eta}$ we have
    $$
    \mathbb{E}\|w^i - w^*\| \leq \left(1 - \frac{1}{\sqrt{n}} \right)^i \|w - w^*\| + C_l
    $$

    $$
    \mathbb{E}\|s^i - 0\| \leq \sqrt{\log(n)} \|w - w^*\| + C_l
    $$

    - (b) If $p_i = \frac{1}{\log(i)}$ and $\alpha = \frac{\eta_i}{\eta}$ is chosen uniformly at random, then with $\alpha = \frac{\eta_i}{\eta}$ we have
      $$
      \mathbb{E}\|w^i - w^*\| \leq \left(1 - \frac{1}{\sqrt{n} \log(n)} \right)^i \|w - w^*\| + C_l
      $$

- (a) SAGA has a linear convergence rate whenever $p_i > 0$ for all $i$.

- (b) SAGA has a faster rate with $p_i$ proportional to $L_i$ and generating a uniform sample.

**Experiment Results**

- Figure: Training objective sub-optimality against effective number of passes for OCR, CONLL-2000, POS.
- Figure: Test error against effective number of passes, for OCR, CONLL-2000, POS. (Dashed lines are stochastic method with sub-optimal step-size.)

**Discussion**

- If memory requirements prohibitive, use mini-batches or SVRG [Johnson & Zhang, 2013].
- Could use $\ell_1$-regularization with proximal versions [Defazio et al., 2014].
- Algorithms applies to any graph structure and approximate inference could be used.
- Could use multi-core computation and distributed parallel implementations.