A fast incremental optimization method with a superlinear rate of convergence

Anton Rodomanov anton.rodomanov@gmail.com
Dmitry Kropotov dmitry.kropotov@gmail.com
Bayesian methods research group (http://bayesgroup.ru), Lomonosov Moscow State University, Moscow, Russia

Motivation

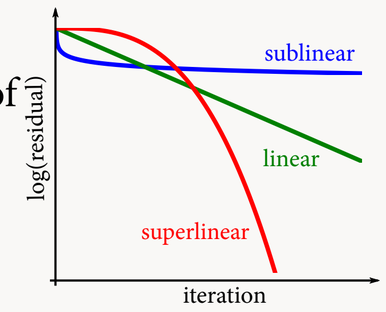
- Minimization of the ℓ_2-regularized average of many functions:

$$\min_{\mathbf{w}\in\mathbb{R}^D} \left[F(\mathbf{w}) := \frac{1}{N}\sum_{i=1}^N f_i(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2\right].$$

- A lot of problems in machine learning have this form.
- Big data setting: N is very large (millions, billions, etc.).

- Incremental/stochastic methods, whose iteration cost does not depend on N, are one of the most effective tools for this task.

- There exist a lot of incremental methods.
- They all have either a sublinear or linear rate of convergence.
- We propose an incremental method with a superlinear rate of convergence.



Assumptions

- All f_i are twice continuously differentiable and convex.
- The Hessians $\nabla^2 f_i$ satisfy the Lipschitz condition:

$$\left\|\nabla^2 f_i(\mathbf{w}) - \nabla^2 f_i(\mathbf{u})\right\|_2 \le M\|\mathbf{w}-\mathbf{u}\|_2, \qquad \forall \mathbf{w},\mathbf{u}\in\mathbb{R}^D.$$

Main idea

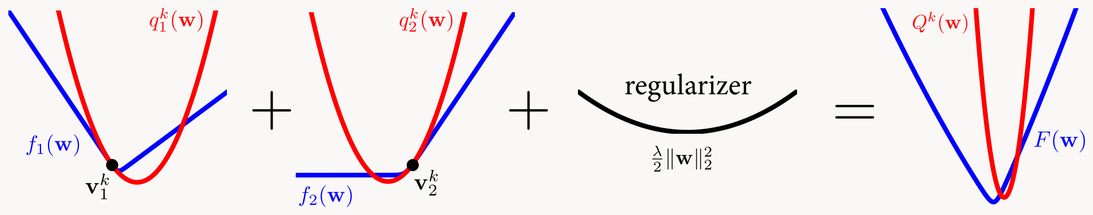
- For each f_i build its own quadratic model:

$$q_i^k(\mathbf{w}) := f_i(\mathbf{v}_i^k) + \nabla f_i(\mathbf{v}_i^k)^\top(\mathbf{w}-\mathbf{v}_i^k) + \frac{1}{2}(\mathbf{w}-\mathbf{v}_i^k)^\top\nabla^2 f_i(\mathbf{v}_i^k)(\mathbf{w}-\mathbf{v}_i^k).$$

- Together they form a quadratic model of F:

$$Q^k(\mathbf{w}) := \frac{1}{N}\sum_{i=1}^N q_i^k(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2.$$

- Step: $\mathbf{w}_{k+1} := \mathbf{w}_k + \alpha_k(\bar{\mathbf{w}}_k - \mathbf{w}_k)$, where $\bar{\mathbf{w}}_k := \operatorname{argmin}_{\mathbf{w}} Q^k(\mathbf{w})$.
- Update only one component q_i^k at each iteration.



The algorithm

Algorithm NIM: a Newton-type incremental method

Require: $\mathbf{w}\in\mathbb{R}^D$: initial point; $K\in\mathbb{N}$: number of iterations.

Initialize: $\mathbf{H}\leftarrow\mathbf{0}^{D\times D}$; $\mathbf{p}\leftarrow\mathbf{0}^D$; $\mathbf{g}\leftarrow\mathbf{0}^D$
  $\mathbf{v}_i\leftarrow$ undefined, $i=1,\dots,N$
for $k=0,1,2,\dots,K-1$ **do**
  Choose an index (cyclic order): $i\leftarrow k \bmod N+1$
  Update the average Hessian, scaled center and gradient:
    $\mathbf{H}\leftarrow\mathbf{H}+(1/N)[\nabla^2 f_i(\mathbf{w})-\nabla^2 f_i(\mathbf{v}_i)]$
    $\mathbf{p}\leftarrow\mathbf{p}+(1/N)[\nabla^2 f_i(\mathbf{w})\mathbf{w}-\nabla^2 f_i(\mathbf{v}_i)\mathbf{v}_i]$
    $\mathbf{g}\leftarrow\mathbf{g}+(1/N)[\nabla f_i(\mathbf{w})-\nabla f_i(\mathbf{v}_i)]$
  Move the ith center: $\mathbf{v}_i\leftarrow\mathbf{w}$
  Find the model's minimum: $\bar{\mathbf{w}}\leftarrow(\mathbf{H}+\lambda\mathbf{I})^{-1}(\mathbf{p}-\mathbf{g})$
  Make a step: $\mathbf{w}\leftarrow\mathbf{w}+\alpha(\bar{\mathbf{w}}-\mathbf{w})$ for some $\alpha>0$
end for
return $\mathbf{w}$

Assume no subtraction is performed when $\mathbf{v}_i$ = undefined.

Theorem (local rate of convergence)

- Let all the centers be initialized close enough to the optimum $\mathbf{w}_*$:

$$\left\|\mathbf{v}_i^0 - \mathbf{w}_*\right\|_2 \le \frac{2\lambda}{M\sqrt{N}}.$$

- Assume the unit step length $\alpha_k\equiv 1$ is used.
- Denote the sequence of iterates of NIM by $\{\mathbf{w}_k\}$.
- Then $\{\mathbf{w}_k\}$ converges to $\mathbf{w}_*$ at an R-superlinear rate:

$$\left\|\mathbf{w}_k-\mathbf{w}_*\right\|_2 \le r_k \qquad \text{and} \qquad \lim_{k\to\infty}\frac{r_{k+1}}{r_k}=0.$$

- Moreover, the rate of convergence is N-step R-quadratic:

$$r_{k+N}\le\frac{M}{2\lambda}r_k^2, \qquad k=2N,2N+1,\dots.$$

Linear models

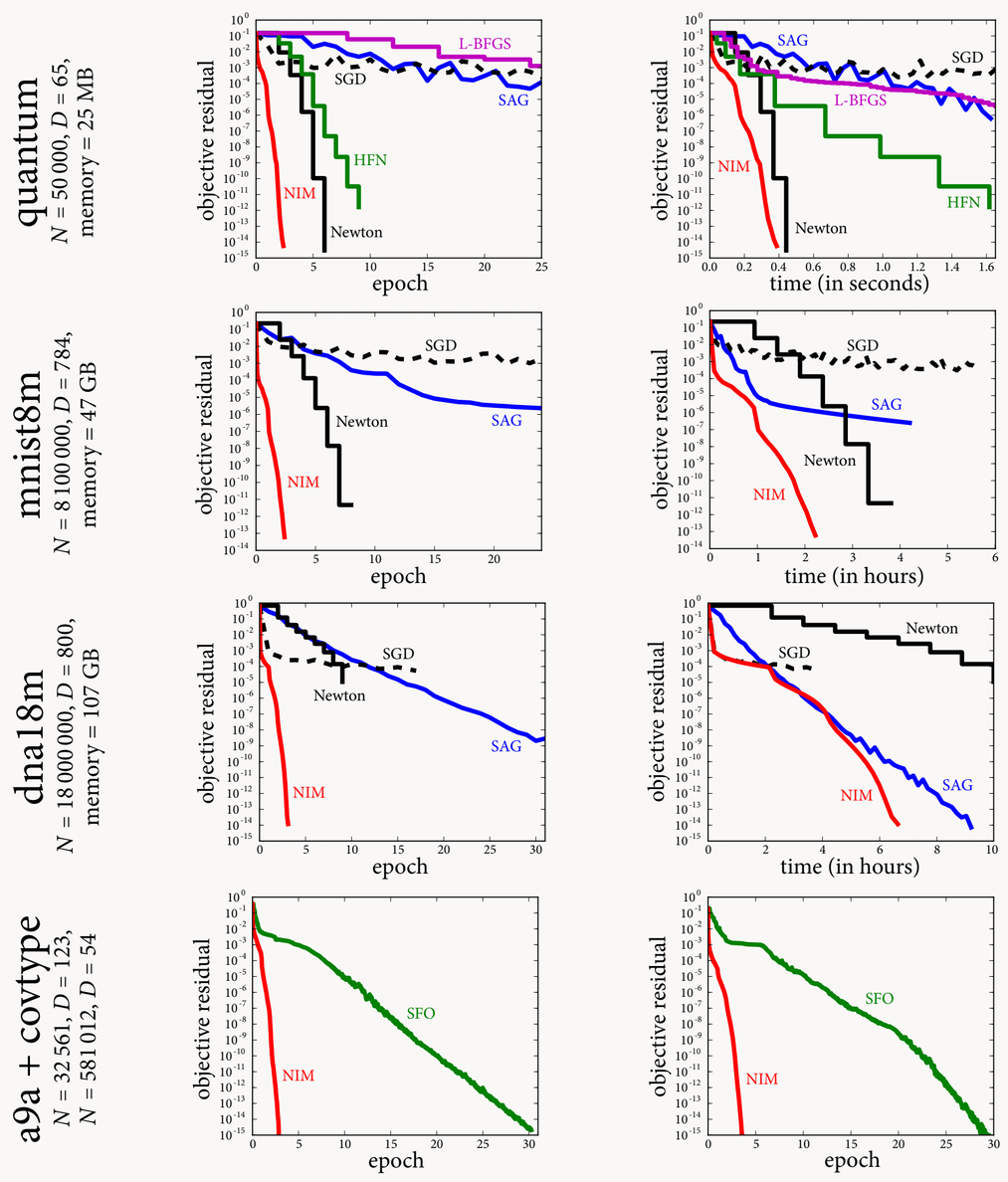
- Linear models: $f_i(\mathbf{w}) := \phi_i(\mathbf{x}_i^\top\mathbf{w})$ for some $\mathbf{x}_i\in\mathbb{R}^D$.
- The gradients and Hessians have a special structure:

$$\nabla f_i(\mathbf{w}) = \phi_i'(\mathbf{x}_i^\top\mathbf{w})\mathbf{x}_i \qquad \text{and} \qquad \nabla^2 f_i(\mathbf{w}) = \phi_i''(\mathbf{x}_i^\top\mathbf{w})\mathbf{x}_i\mathbf{x}_i^\top.$$

- Instead of $\mathbf{v}_i^k$ store the corresponding dot product $\mu_i^k := \mathbf{x}_i^\top\mathbf{v}_i^k$.
- Work directly with $\mathbf{B}_k := (\mathbf{H}_k+\lambda\mathbf{I})^{-1}$ using rank-1 updates.

Method	Iteration cost	Memory	Rate of convergence	
			In iterations	In epochs
SGD	$O(D)$	$O(D)$	Sublinear	Sublinear
SAG	$O(D)$	$O(N+D)$	Linear	Linear
NIM	$O(D^2)$	$O(N+D^2)$	Superlinear	Quadratic

Experiments (logistic regression)



Future work

- Global convergence: proof + line search procedure.

References

[1] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951.
[2] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv*, 2013.
[3] J. Sohl-Dickstein, B. Poole and S. Ganguli. Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods. *31th International Conference on Machine Learning*, 2014.