A Newton-type Incremental Method with a Superlinear Convergence Rate

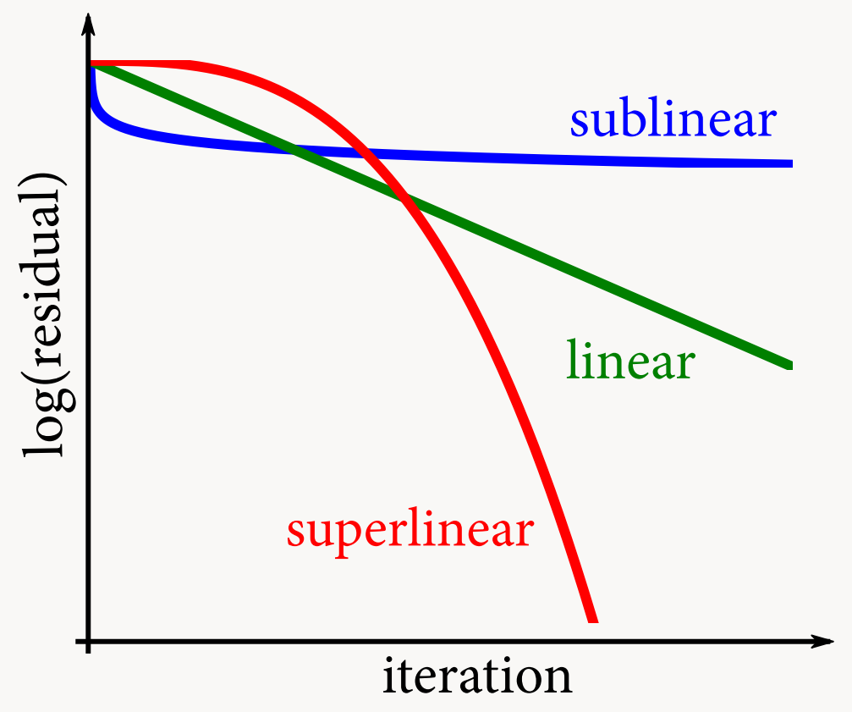
Anton Rodomanov   Dmitry Kropotov
anton.rodomanov@gmail.com   dmitry.kropotov@gmail.com
Bayesian methods research group (http://bayesgroup.ru), Higher School of Economics and Lomonosov Moscow State University, Moscow, Russia

Motivation

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

$$\min_{x \in \mathbf{R}^d} \left[f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \frac{\mu}{2}\|x\|_2^2 \right].$$

- 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 among the most effective methods for this task.
- There exist a lot of incremental methods.
- They all have either a sublinear or linear convergence rate.
- We propose an incremental method with a superlinear convergence rate.



Assumptions

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

$$\|\nabla f_i(x) - \nabla f_i(y)\|_2 \le L_f \|x - y\|_2,$$
$$\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_2 \le M \|x - y\|_2$$

for all $x, y \in \mathbf{R}^d$.

The algorithm

```
Algorithm NIM: a Newton-type Incremental Method
Require: $x \in \mathbf{R}^d$: initial point; $K \in \mathbf{N}$: number of iterations.
Initialize: $H \leftarrow 0^{d\times d}$; $u \leftarrow 0^d$; $g \leftarrow 0^d$; $v_i \leftarrow$ 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:
        $H \leftarrow H + (1/n)[\nabla^2 f_i(x) - \nabla^2 f_i(v_i)]$
        $u \leftarrow u + (1/n)[\nabla^2 f_i(x)x - \nabla^2 f_i(v_i)v_i]$
        $g \leftarrow g + (1/n)[\nabla f_i(x) - \nabla f_i(v_i)]$
    Move the $i$th center: $v_i \leftarrow x$
    Find the model's minimum: $\bar{x} \leftarrow (H + \mu I)^{-1}(u - g)$
    Make a step: $x \leftarrow x + \alpha(\bar{x} - x)$ for some $\alpha > 0$ (usually $\alpha = 1$)
end for
return $x$
```

Assume no subtraction is performed when v_i = undefined.

Main idea

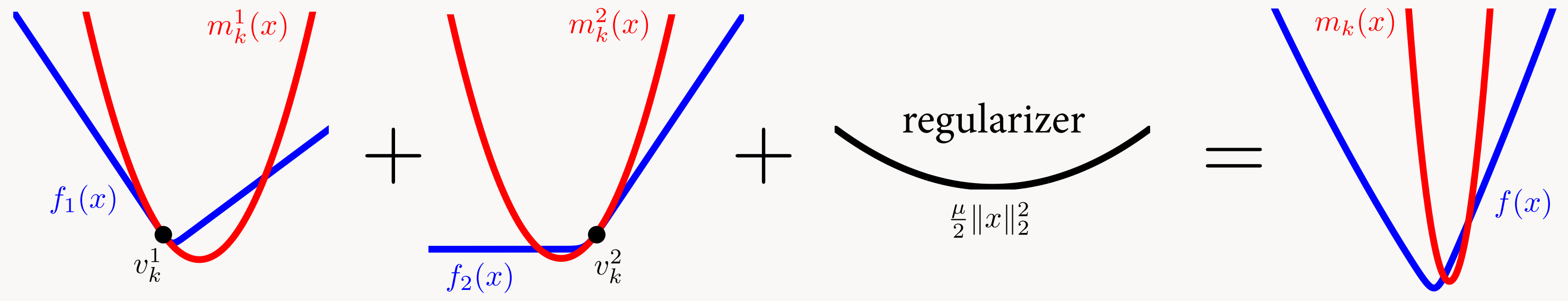
- For each f_i build its own quadratic model:

$$m_k^i(x) := f_i(v_k^i) + \nabla f_i(v_k^i)^\top (x - v_k^i) + \frac{1}{2}(x - v_k^i)^\top \nabla^2 f_i(v_k^i)(x - v_k^i).$$

- Together they form a quadratic model of f:

$$m_k(x) := \frac{1}{n}\sum_{i=1}^{n} m_k^i(x) + \frac{\mu}{2}\|x\|_2^2.$$

- Step: $x_{k+1} := x_k + \alpha_k(\bar{x}_k - x_k)$, where $\bar{x}_k := \operatorname{argmin}_x m_k(x)$.
- Update only one component m_k^i at each iteration.



Theorem (local convergence)

- Let all the centers be initialized close enough to the optimum x^*:

$$\|v_0^i - x^*\|_2 \le \frac{2\mu}{M\sqrt{n}}, \qquad i = 1, \dots, n.$$

- Assume the unit step length $\alpha_k \equiv 1$ is used.
- Then $\{x_k\}$ converges to x^* at an R-superlinear rate:

$$\|x_k - x^*\|_2 \le r_k \qquad \text{and} \qquad \lim_{k\to\infty} \frac{r_{k+1}}{r_k} = 0.$$

- More precisely, the convergence rate of $\{x_k\}$ is n-step R-quadratic:

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

Theorem (global convergence)

- Denote the condition number of f as $\kappa := (L_f + \mu)/\mu$.
- Assume a constant step: $\alpha_k \equiv \alpha < \bar{\alpha} := 2\kappa^{-3}(1 + 19\kappa(n-1))^{-1}$.
- Then, for any initialization, $\{x_k\}$ converges to x^* at an R-linear rate:

$$\|x_k - x^*\|_2 \le \sqrt{\kappa} \cdot c^{k/2} \|x_0 - x^*\|_2,$$

where $c := h^{1/(1+2(n-1))}$ for $h := 1 - 2\kappa^{-1}\alpha + \kappa^2(1 + 19\kappa(n-1))\alpha^2$.

Theoretical comparison

Method	Iteration cost	Memory	Convergence rate	
			In iterations	In epochs
SGD [1]	$O(D)$	$O(D)$	Sublinear	Sublinear
SAG [2]	$O(D)$	$O(ND)$	Linear	Linear
SFO [3]	$O(ND)$	$O(ND)$	Linear?	Linear?
NIM	$O(D^3)$	$O(ND + D^2)$	Superlinear	Quadratic

Linear models

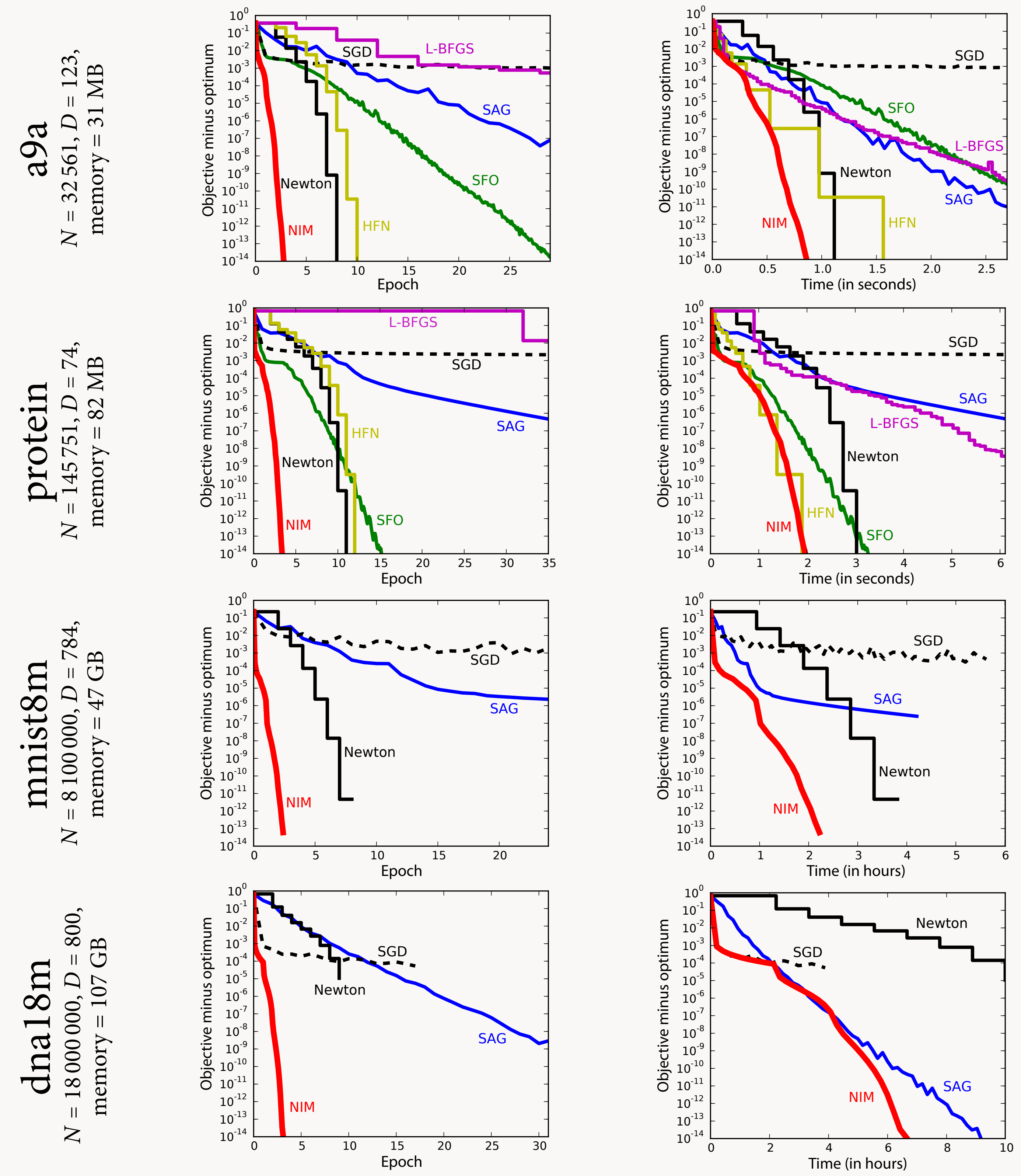
- Linear models: $f_i(x) := \phi_i(a_i^\top x)$ for some $a_i \in \mathbf{R}^d$.
- The gradients and Hessians have a special structure:

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

- Instead of v_k^i store the corresponding dot products $\nu_k^i := a_i^\top v_k^i$.
- Work directly with $B_k := (H_k + \mu I)^{-1}$ using rank-1 updates.

Method	Iteration cost	Memory	Convergence rate	
			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)



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.