Proximal Incremental Newton Method

Anton Rodomanov

Higher School of Economics

26 February 2016

Seminar on Bayesian Methods in Machine Learning, Moscow, Russia

Problem formulation

- We consider the following strongly convex optimization problem:

$$\min_{x \in \mathbf{R}^d} \left[\phi(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x) \right],$$

where

  - f_i: differentiable convex functions;
  - h: convex and simple (possibly non-differentiable).

Example (regularized empirical risk minimization)

We want to fit a parametric model to the data.

- x: parameters of the model;
- $f_i(x)$: error of the model on the ith training sample;
- $h(x)$: regularizer, e.g. $h(x) = \lambda \|x\|_1$.

Problem formulation

- We consider the following strongly convex optimization problem:

$$\min_{x \in \mathbf{R}^d} \left[\phi(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x) \right],$$

where

- f_i: differentiable convex functions;
- h: convex and simple (possibly non-differentiable).

Example (constrained empirical risk minimization)

Let $Q \subseteq \mathbf{R}^d$ be a convex set and h be the indicator function of Q:

$$h(x) = \begin{cases} 0, & x \in Q, \\ +\infty, & x \notin Q. \end{cases}$$

Then the unconstrained minimization of ϕ is equivalent to

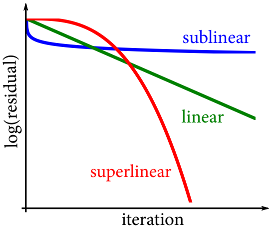
$$\min_{x \in Q} \quad \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

Motivation

- n is very large $\Rightarrow$ interested in methods whose iteration cost is independent of n.
- These methods are called **incremental methods** [Bertsekas, 2011]:
  - **Purely stochastic**: $x_{k+1} = \text{prox}_h(x_k - \alpha_k B_k \nabla f_{i_k}(x_k))$.
    - SGD [Robbins-Monro, 1951], oLBFGS [Schraudolph et al., 2007], AdaGrad [Duchi et al., 2011], SQN [Byrd et al., 2014], Adam [Kingma, 2014] etc.
    - Convergence rate: sublinear, usually $O(1/k)$.
  - **Purely incremental**.
    - IAG [Blatt et al., 2007], SAG [Schmidt et al., 2013], SVRG [Johnson & Zhang, 2013], FINITO [Defazio et al., 2014b], SAGA [Defazio et al., 2014a], MISO [Mairal, 2015] etc.
    - Reducing variance: use an estimate $g_k \approx \nabla f(x_k)$ whose variance tends to zero as $x_k \to x^*$.
    - Convergence rate: linear, $O(c^k)$.
- **Goal**: an incremental method with a superlinear convergence rate.

Convergence rates

- Consider some iterative optimization method for solving $\min_x \phi(x)$.
- It generates the sequence $\{x_k\}$ such that $\phi(x_k) \to \phi(x^*)$ as $k \to \infty$.
- Define the sequence of residuals $\{r_k\}$ such that $r_k \geq 0$ and $r_k \to 0$, e.g. $r_k := \|x_k - x^*\|$ or $r_k := \phi(x_k) - \phi(x^*)$.
- **Convergence rates**:
  - **Linear**: $r_{k+1} \leq c r_k$, where $0 < c < 1$.
  - **Sublinear**: $r_{k+1} \leq c_k r_k$, where $c_k \uparrow 1$.
  - **Superlinear**: $r_{k+1} \leq c_k r_k$, where $c_k \downarrow 0$.



Newton method

Consider the following problem:

$$\min_x f(x),$$

where f is twice continuously differentiable and strongly convex.

- Let x_k be the current iterate and $H_k := \nabla^2 f(x_k) \succ 0$.
- Consider the second-order Taylor approximation of f around x_k:
$$m_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle.$$
- Find the minimum of the model: $\bar{x}_k := \operatorname{argmin}_x m_k(x)$,
$$\bar{x}_k = x_k - H_k^{-1} \nabla f(x_k).$$
- **Newton step**: $x_{k+1} := x_k + \alpha_k (\bar{x}_k - x_k)$,
$$x_{k+1} = x_k - \alpha_k H_k^{-1} \nabla f(x_k),$$
where $\alpha_k > 0$ is the step length:
    - $\alpha_k \equiv 1$: pure Newton method;
    - $\alpha_k \not\equiv 1$: damped Newton method.

Convergence of Newton method

Theorem (local convergence rate)

Suppose

- f *is strongly convex with constant* $\mu_f > 0$;
- $\nabla^2 f$ *is Lipschitz-continuous with constant* $M_f > 0$.
- **the initial point** x_0 **is close enough to** x^*:
$$\|x_0 - x^*\| \leq \frac{\mu_f}{2M_f}.$$

Then the sequence $\{x_k\}_{k\geq 0}$, *generated by the* **pure Newton** *method* $(\alpha_k \equiv 1)$, *converges to* x^* *at a* *superlinear* *(quadratic) rate:*
$$\|x_{k+1} - x^*\| \leq \frac{M_f}{\mu_f} \|x_k - x^*\|^2 .$$

Convergence of Newton method

Theorem (global convergence rate)

Suppose

- f *is strongly convex with constant* $\mu_f > 0$;
- ∇f *is Lipschitz-continuous with constant* $L_f > 0$.

Then, **for any initial point** x_0, *the* **damped Newton** *method with* $\alpha_k = \mu_f / L_f$ *generates a sequence* $\{x_k\}$ *such that* $\{f(x_k)\}$ *converges to* $f(x^*)$ *at a linear rate:*

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu_f^2}{L_f^2}\right) [f(x_k) - f(x^*)].$$

Incremental Newton Method

Consider the sum-of-functions problem:

$$\min_x \left[f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right],$$

where each f_i is twice continuously differentiable and strongly convex.

- Build the second-order Taylor model of each f_i:

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

- Model of the full function f: $m_k(x) := (1/n) \sum_{i=1}^n m_k^i(x)$.
- **Iteration** k:
  - Make a step: $x_{k+1} := x_k + \alpha_k (\bar{x}_k - x_k)$, where $\bar{x}_k := \operatorname{argmin}_x m_k(x)$.
  - Update the model: choose $i_k \in \{1, \ldots, n\}$ and change only one component of the model:

$$v_{k+1}^i := \begin{cases} x_{k+1}, & i = i_k, \\ v_k^i, & \text{otherwise.} \end{cases}$$

Efficient update of the model

Note that m_k is a quadratic function

$$m_k(x) = \frac{1}{2}\langle H_k x, x\rangle + \langle g_k - u_k, x\rangle + \mathrm{const},$$

where

$$H_k := \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(v_k^i), \quad g_k := \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(v_k^i), \quad u_k := \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(v_k^i) v_k^i.$$

- Since we update only one component at every iteration,

$$H_{k+1} = H_k + \frac{1}{n}[\nabla^2 f_i(v_{k+1}^i) - \nabla^2 f_i(v_k^i)],$$
$$g_{k+1} = g_k + \frac{1}{n}[\nabla f_i(v_{k+1}^i) - \nabla f_i(v_k^i)],$$
$$u_{k+1} = u_k + \frac{1}{n}[\nabla^2 f_i(v_{k+1}^i) v_{k+1}^i - \nabla^2 f_i(v_k^i) v_k^i].$$

- If we store v_k^i in memory, the cost of the update is independent of n.
- Cost of finding $\bar{x}_k := H_k^{-1}(u_k - g_k)$ is also independent of n.

The algorithm

Input: $x_0, \ldots, x_{n-1} \in \mathbf{R}^d$: initial points; $\alpha > 0$: step length.

Initialize model:

$$H \leftarrow (1/n) \sum_{i=1}^{n} \nabla^2 f_i(x_{i-1})$$
$$g \leftarrow (1/n) \sum_{i=1}^{n} \nabla f_i(x_{i-1})$$
$$u \leftarrow (1/n) \sum_{i=1}^{n} \nabla^2 f_i(x_{i-1}) x_{i-1}$$
$$v_i \leftarrow x_{i-1}, \ \ i = 1, \ldots, n$$

for $k \geq n-1$ **do**

    Minimize model: $\bar{x} \leftarrow H^{-1}(u - g)$

    Make a step: $x_{k+1} \leftarrow x_k + \alpha(\bar{x} - x_k)$

    Update model:

$$i \leftarrow (k+1) \bmod n + 1$$
$$H \leftarrow H + (1/n)[\nabla^2 f_i(x_{k+1}) - \nabla^2 f_i(v_i)]$$
$$g \leftarrow g + (1/n)[\nabla f_i(x_{k+1}) - \nabla f_i(v_i)]$$
$$u \leftarrow u + (1/n)[\nabla^2 f_i(x_{k+1}) x_{k+1} - \nabla^2 f_i(v_i) v_i]$$
$$v_i \leftarrow x_{k+1}$$

end for

Convergence of incremental Newton method

Theorem (Local convergence rate)

Suppose **all the initial points** $x_0, \dots, x_{n-1}$ **are close enough to** x^**:*

$$\|x_i - x^*\| \leq \frac{\mu_f}{2M_f}, \qquad i = 1, \dots, n.$$

Then the sequence $\{x_k\}$ *generated by the* **pure incremental Newton method** *(*$\alpha_k \equiv 1$*), converges to* x^* *at a superlinear rate:*

$$\begin{aligned} \|x_k - x^*\| &\leq z_k, & k &\geq 0, \\ z_{k+1} &\leq c_k z_k, & k &\geq n, \end{aligned}$$

where

$$c_k := \left(1 - \frac{3}{4n}\right)^{2^{\lceil k/n \rceil - 1}}$$

More precisely, the converge rate is n-step quadratic:

$$z_{k+n} \leq \frac{M_f}{\mu_f} z_k^2, \qquad k \geq 0.$$

Convergence of incremental Newton method

Theorem (Global convergence rate)

Denote the condition number of f as $\kappa_f := L_f/\mu_f$. Then, **for any initial points** $x_0, \dots, x_{n-1}$, *the* **damped incremental Newton method** *with* $\alpha_k = \kappa_f^{-3}(1 + 19\kappa_f(n-1))^{-1}$ *generates a sequence* $\{x_k\}$ *such that* $\{f(x_k)\}$ *converges to* $f(x^*)$ *at a linear rate:*

$$f(x_k) - f(x^*) \leq c^k [f(x_0) - f(x^*)],$$

where

$$c := \left(1 - \kappa_f^{-4}(1 + 19\kappa_f(n-1))^{-1}\right)^{\frac{1}{1+2(n-1)}}$$

Efficient model minimization for linear models

The minimum of the model m_k is given by
$$\bar{x}_k = H_k^{-1}(u_k - g_k).$$

- Consider **linear models**: $f_i(x) := \phi_i(\langle a_i, x\rangle)$ for some $a_i \in \mathbf{R}^d$.
- Denote $\nu_k^i := \langle a_i, v_k^i\rangle$. Then the model update can be written as
$$H_{k+1} = H_k + \frac{1}{n}[\phi_i''(\nu_{k+1}^i) - \phi_i''(\nu_k^i)]a_i a_i^\top$$
$$g_{k+1} = g_k + \frac{1}{n}[\phi_i'(\nu_{k+1}^i) - \phi_i'(\nu_k^i)]a_i,$$
$$u_{k+1} = u_k + \frac{1}{n}[\phi_i''(\nu_{k+1}^i)\nu_{k+1}^i - \phi_i''(\nu_k^i)\nu_k^i]a_i.$$
- Memory requirement reduces from $O(nd + d^2)$ to $O(n + d^2)$.
- **Rank-1 update** $\Rightarrow$ apply Sherman-Morrison formula to $B_k := H_k^{-1}$:
$$B_{k+1} = B_k - \frac{\delta_k B_k a_i a_i^\top B_k}{n + \delta_k \langle B_k a_i, a_i\rangle}, \qquad \delta_k := \phi_i''(\nu_{k+1}^i) - \phi_i''(\nu_k^i).$$
- Iteration cost reduces from $O(d^3)$ to $O(d^2)$.

Proximal gradient method

Consider the minimization of a **composite function**:

$$\min_{x \in \mathbf{R}^d} \; \phi(x) := f(x) + h(x),$$

where

- f is differentiable and convex;
- h is convex and simple (possibly non-differentiable).

- **Proximal gradient method**:

  $$x_{k+1} = \operatorname{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)),$$

  where $\alpha_k > 0$ is the step length and prox is the **proximal operator**:

  $$\operatorname{prox}_{\alpha h}(x) := \operatorname*{argmin}_y \left[\alpha h(y) + \frac{1}{2} \|y - x\|^2 \right].$$

- This is equivalent to minimizing a separable quadratic model:

  $$x_{k+1} = \operatorname*{argmin}_x \left[f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right]$$

- Assumption "h is simple" means we can efficiently compute $\operatorname{prox}_{\alpha h}(\cdot)$.

Convergence of proximal gradient method

All the convergence results of the standard gradient method are retained as if there were no $h(x)$. For example:

Theorem

Assume

- f *is strongly convex with constant* $\mu_f > 0$*;*
- ∇f *is Lipschitz-continuous with constant* $L_f > 0$*.*

Then, for any initial x_0*, the proximal gradient method converges to* x^* *at a linear rate:*

$$\|x_{k+1} - x^*\| \leq \left(\frac{\kappa_f - 1}{\kappa_f + 1} \right) \|x_k - x^*\| .$$

Evaluating proximal mapping: examples

- (**ℓ_1-norm regularization**) $h(x) := \|x\|_1$:

$$[\mathrm{prox}_{\alpha h}(x)]_i = \begin{cases} x_i - \alpha, & x_i > \alpha, \\ 0, & |x_i| \le \alpha, \\ x_i + \alpha, & x_i < -\alpha. \end{cases}$$

- (**Indicator of a convex set**) $h(x) := I_Q(x)$:

$$h(x) = \operatorname*{argmin}_{y \in Q} \frac{1}{2} \|y - x\|^2 ,$$

  i.e. proximal operator generalizes the projection operator.

- (**Elastic net regularization**) $h(x) := \|x\|_1 + (\gamma/2) \|x\|_2^2$:

$$\mathrm{prox}_{\alpha h}(x) = \left(\frac{1}{1 + \alpha\gamma} \right) \mathrm{prox}_{\alpha \|\cdot\|_1}(x).$$

Proximal Newton method

Minimization of a composite function:

$$\min_{x \in \mathbf{R}^d} \ \phi(x) := f(x) + h(x).$$

- Let x_k be the current iterate and $H_k := \nabla^2 f(x_k) \succ 0$.
- Build a model of ϕ around x_k:

$$m_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle H_k(x - x_k), x - x_k \rangle + h(x).$$

- Find the minimum of the model: $\bar{x}_k := \operatorname{argmin} m_k(x)$,

$$\bar{x}_k = \operatorname{prox}_h^{H_k}(x_k - H_k^{-1} \nabla f(x_k)),$$

  where prox is the **scaled proximal operator**:

$$\operatorname{prox}_h^H(x) := \operatorname*{argmin}_y \left[h(y) + \frac{1}{2}\|y - x\|_H^2 \right].$$

- **Proximal Newton step**:

$$x_{k+1} := x_k + \alpha_k(\bar{x}_k - x_k).$$

Evaluating the scaled proximal mapping

Scaled proximal mapping:

$$\mathrm{prox}_h^H(x) := \operatorname*{argmin}_y \left[h(y) + \frac{1}{2} \|y - x\|_H^2 \right].$$

- Cannot be computed analytically even if h is separable (e.g. ℓ_1-norm) $\Rightarrow$ need to use an auxiliary optimization method for this subproblem.
- We need to minimize a composite function
$$\Phi(y) := h(y) + F(y).$$
- Define **composite gradient mapping** of a function f:
$$g_\alpha^f(y) := \frac{1}{\alpha}(y - \mathrm{prox}_{\alpha h}(y - \alpha \nabla f(y))),$$
where $\alpha > 0$ is some step length. If $h \equiv 0$, then $g_\alpha^f(y) = \nabla f(y)$.
- **Termination criterion** for the inner method: stop at y if
$$\left\| g_\alpha^F(y) \right\| \le \min\{1, \Delta_k^\gamma\} \Delta_k, \qquad \Delta_k := \left\| g_1^f(x_k) \right\|$$
- Possible inner method: Fast Gradient Method [Nesterov, 2013].

Convergence of proximal Newton method

Due to the special termination criterion for the inner method:

- We do not spend much effort on solving the subproblem accurately at early iterations of Newton method $\Rightarrow$ iteration complexity decreases.
- We do not loose anything in convergence rates:
  - pure Newton $\Rightarrow$ superlinear convergence;
  - damped Newton $\Rightarrow$ linear convergence.

Incremental proximal Newton method

$$\min_x \left[\phi(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x) \right].$$

- Incorporate h into the model:

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

$$m_k(x) := \frac{1}{n} \sum_{i=1}^{n} m_k^i(x) + h(x).$$

- Everything is the same. Now $\bar{x}_k$ becomes $\bar{x}_k = \mathrm{prox}_h^{H_k}[H_k^{-1}(u_k - g_k)]$.
- **Termination criterion** for the inner method: stop at y if

$$\left\| g_\alpha^F(y) \right\| \leq \min\{1, \Delta_k^\gamma\} \Delta_k,$$

where Δ_k is the **incremental composite gradient mapping**:

$$\Delta_k := \| \bar{v}_k - \mathrm{prox}_h(\bar{v}_k - g_k) \|, \qquad \bar{v}_k = \frac{1}{n} \sum_{i=1}^{n} v_k^i.$$

Convergence of incremental proximal Newton method

Inexact solution of the subproblem does not kill superlinear convergence:

Theorem

Suppose all the initial points $x_0, \dots, x_{n-1}$ are close enough to x^*:

$$\|x_i - x^*\| \le \min\left\{ \frac{\mu_f}{2M_f},\ \left(\frac{\mu_f^3}{128(2 + L_f)^{5+2\gamma}} \right)^{1/(2\gamma)} \right\}.$$

Then the sequence $\{x_k\}$, generated by the pure incremental proximal Newton method ($\alpha_k \equiv 1$), converges to x^* at a superlinear rate:

$$\|x_k - x^*\| \le z_k, \qquad k \ge 0$$

$$z_{k+1} \le c_k z_k, \qquad k \ge n,$$

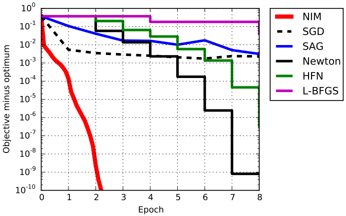
where

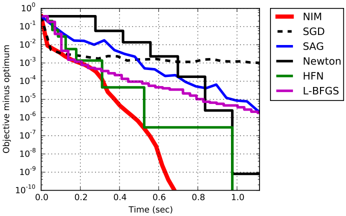
$$c_k := \left(1 - \frac{7}{16n}\right)^{(1+\gamma)^{\lceil k/n \rceil}/2}$$

Analogously, there is a theorem about global linear convergence.

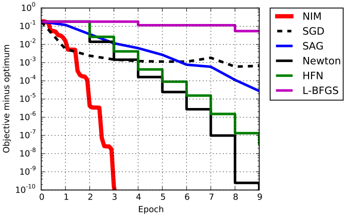
Experimental results: ℓ_2-regularized logistic regression

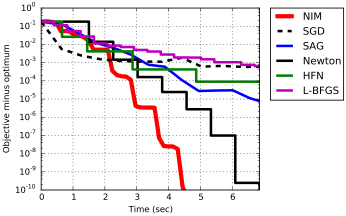
- a9a ($n = 32\,000$, $d = 123$)





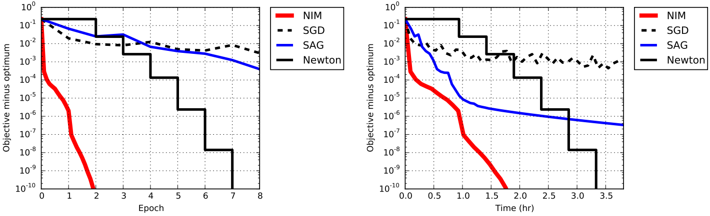
- covtype ($n = 500\,000$, $d = 54$)



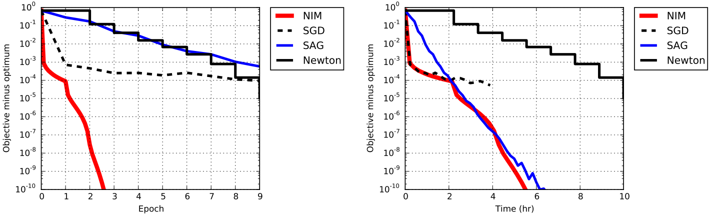


Experimental results: ℓ_2-regularized logistic regression

- mnist8m ($n = 8\,000\,000$, $d = 784$)

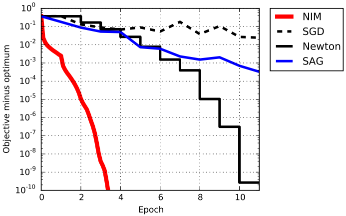


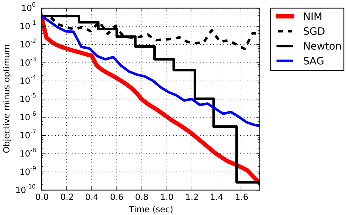
- dna18m ($n = 18\,000\,000$, $d = 800$)



Experimental results: ℓ_1-regularized logistic regression

- a9a $(n = 32\,000,\ d = 123)$





- covtype $(n = 500\,000,\ d = 54)$