Efficient Optimization in Structured Learning

Xiaocheng Tang with
Katya Scheinberg

Q&A

- can we apply L-BFGS to non-smooth function?
- can we combine L-BFGS with Randomized Coordinate Descent?
- is it faster than ISTA/FISTA?
- is it faster than L-BFGS with ISTA/FISTA?
- what is the complexity for each RCD step?
- how many RCD steps should we run per iteration?
- how many RCD steps do we need to achieve ϵ-accuracy?

Objective

- consider minimizing the following composite function:

$$\min_{x \in \mathbb{R}^n} F(x) \equiv f(x) + g(x)$$

- for example, sparse optimization
  - Classification – Sparse Logistic Regression (SLR)

$$f(w) = \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-y_i \cdot w^T x_i)), \quad g(w) = \lambda \|w\|_1, \quad w \in \mathbb{R}^p$$

    training set $\{(x_i, y_i)\}_{i=1}^N \in (\mathbb{R}^p \times \{-1, 1\})$
  - Graphical model – Sparse Inverse Covariance Selection (SICS)

$$f(X) = -\log\det X + \mathbf{tr}(SX), \quad g(X) = \lambda \|X\|_1, \quad X \in \mathbb{S}_{++}^p$$

    low rank sample covariance matrix $S \in \mathbb{S}_+^p$ – more observations than number of random variables.

- and many others, e.g., elastic net, group lasso, matrix completion (with nuclear norm), dictionary learning (with hierarchical norm), etc.

Outer loop

$Q(H, u, v)$

```
For k = 1,2,…:
    Construct local approximation
```

Inner loop

```
    For j = 1,2,…:
        Minimize local approximation
Update variables
```

Outer loop

$Q(H, u, v)$

```
For k = 1,2,…:

    Construct local approximation
```

what if the local approximation is bad?

Inner loop

```
    For j = 1,2,…:

        Minimize local approximation

Update variables
```

Outer loop

$Q(H, u, v)$

Check Sufficient Decrease

Inner loop

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…:
      Minimize local approximation
  Update variables
```

Outer loop

$Q(H, u, v)$

Check Sufficient Decrease

Inner loop

Inexact Solver

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…:
      Minimize local approximation
  Update variables
```

Outer loop

Inner loop

Inexact Solver

$Q(H, u, v)$

Check Sufficient Decrease

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…:
      Minimize local approximation
  Update variables
```

until when?

Outer loop

$Q(H, u, v)$

Check Sufficient Decrease

Inner loop

Inexact Solver

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…,k:
      Minimize local approximation
  Update variables
```

Outer loop

$Q(H, u, v)$

Check Sufficient Decrease

Inner loop

Inexact Solver

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…,l(k):
      Minimize local approximation
  Update variables
```

Outer loop

$Q(H, u, v)$

Check Sufficient Decrease

```
For k = 1,2,…:
   Construct local approximation
   While di                    ot good:
      Update local approximation
```

What is l(k)?

Inner loop

Inexact Solver

```
For j = 1,2,…,l(k):
   Minimize local approximation
```

```
Update variables
```

Basics

- Objective $F(x)$ and Local approximation $Q(H, v, u)$

$$F(x) = f(x) + g(x),$$

$$Q(H, v, u) = f(u) + \langle \nabla f(u), v - u \rangle + \frac{1}{2} \langle v - u, H(v - u) \rangle + g(v).$$

- Exact minimizer $p_H(u)$ and Inexact minimizer $p_{H,\phi}(u)$

$$p_H(u) = \arg\min_v Q(H, v, u),$$

$$Q(H, p_{H,\phi}(u), u) \leq Q(H, u, u) = F(u),$$

$$\text{and } Q(H, p_{H,\phi}(u), u) \leq Q(H, p_H(u), u) + \phi$$

- Sufficient decrease condition

$$x^{k+1} := p_{H_k}(x^k) \text{ or } p_{H_k,\phi_k}(x^k)$$

$$F(x^{k+1}) - F(x^k) \leq \rho(Q(H_k, x^{k+1}, x^k) - F(x^k))$$

- Hessian or Hessian approximation G (and H)

$$H \leftarrow \frac{1}{\mu} I + G$$

Outer loop

$Q(G_k, \nabla f_k, \cdot, x^k)$

Check Sufficient Decrease

$\mu_k \leftarrow \mu_k/2$

Inner loop

$p_{H_k,\phi_k}(x^k)$

$x^{k+1} \leftarrow p_{H_k,\phi_k}(x^k)$

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…,l(k):
      Minimize local approximation
  Update variables
```

Assumptions

- **Existence** The set of optimal solutions, X^*, is nonempty; x^* is any element of X^*.
- **Bounded Level Set** The effective domain of F is defined as $\mathbf{dom}(F) := \{x \in \mathbb{R}^n : F(x) < \infty\}$, and the level set of F at point $x \in \mathbf{dom}(F)$ is defined by

$$\mathcal{X}_F(x) := \{y \in \mathbf{dom}(F) : F(y) \leq F(x)\}.$$

  Without loss of generality, we restrict our discussions below to the level set $\mathcal{X}_0 := \mathcal{X}_F(x^0)$ given by some $x^0 \in \mathbf{dom}(F)$, e.g., the initial iterate.
- **Lipschitz continuity** g is convex and Lipschitz continuous with constant L_g for all $x, y \in \mathcal{X}_0$:

$$g(x) - g(y) \leq L_g \|x - y\|,$$

- **Bounded H** There exists positive constants M and σ such that for all $k \geq 0$, at the k-th iteration:

$$\sigma I \preceq \sigma_k I \preceq H_k \preceq M_k I \preceq M I$$

- There exists a positive constant $D_{\mathcal{X}_0}$ such that for all iterates $\{x^k\}$:

$$\sup_{x^* \in X^*} \|x^k - x^*\| \leq D_{\mathcal{X}_0}$$

Exact Case

Theorem

Given $x^0 \in \mathbb{R}^n$, let the sequence $\{x^k\}$ be generated such that for all $k \geq 0$, $x^{k+1} \leftarrow p_{H_k}(x^k)$ with sufficient decrease held at $p_{H_k}(x^k)$. Then the sequence $\{x^k\}$ satisfy

$$\Delta F_k := F(x^k) - F^* \leq \frac{2M^2(D_{\mathcal{X}_0} M + 2L_g)^2}{\rho\sigma^3} \frac{1}{k}.$$

Remark

- if $H_k = L(f)I$ for all k, as in standard proximal gradient methods, where $L(f)$ is the Lipschitz constant of $\nabla f(x)$, then the bound becomes

$$F(x^k) - F^* \leq \frac{2(D_{\mathcal{X}_0} L(f) + 2L_g)^2}{\rho L(f)} \frac{1}{k} \approx \frac{2D_{\mathcal{X}_0}^2 L(f)}{k},$$

- if $L_g \ll D_{\mathcal{X}_0} L(f)$. This bound is similar to $\frac{2\|x^0 - x^*\|^2 L(f)}{k}$ established for proximal gradient methods, assuming that $D_{\mathcal{X}_0}$ is comparable to $\|x^0 - x^*\|$.

Exact Case

Theorem

Given $x^0 \in \mathbb{R}^n$, let the sequence $\{x^k\}$ be generated such that for all $k \geq 0$, $x^{k+1} \leftarrow p_{H_k}(x^k)$ with sufficient decrease held at $p_{H_k}(x^k)$. Then the sequence $\{x^k\}$ satisfy

$$\Delta F_k := F(x^k) - F^* \leq \frac{2M^2(D_{\mathcal{X}_0} M + 2L_g)^2}{\rho\sigma^3} \frac{1}{k}.$$

Proof.

- $F(x^k) - F(x^{k+1}) \geq c\Delta F_k^2$,
  - $\Delta F_k - \Delta F_{k+1} = F(x^k) - F(x^{k+1}) \geq c\Delta F_k^2$
  - $\frac{1}{\Delta F_k} - \frac{1}{\Delta F_{k+1}} \geq c\frac{\Delta F_k}{\Delta F_{k+1}} \geq c$
  - $\frac{1}{\Delta F_k} \geq kc + \frac{1}{\Delta F_0} \geq kc$

- Nesterov [2004], Nesterov and Polyak [2006], Cartis et al. [2012]

Exact Case

Theorem

Given $x^0 \in \mathbb{R}^n$, let the sequence $\{x^k\}$ be generated such that for all $k \geq 0$, $x^{k+1} \leftarrow p_{H_k}(x^k)$ with sufficient decrease held at $p_{H_k}(x^k)$. Then the sequence $\{x^k\}$ satisfy

$$\Delta F_k := F(x^k) - F^* \leq \frac{2M^2(D_{\mathcal{X}_0} M + 2L_g)^2}{\rho\sigma^3} \frac{1}{k}.$$

Proof.

- $F(x^k) - F(x^{k+1}) \geq c\Delta F_k^2$,
- $\Delta F_k \leq c_1 \|\nabla f(x^k) + \gamma_g^{k+1}\|$,
- $F(x^k) - F(x^{k+1}) \geq c_2 \|\nabla f(x^k) + \gamma_g^{k+1}\|^2$.

Exact Case

Theorem

Given $x^0 \in \mathbb{R}^n$, let the sequence $\{x^k\}$ be generated such that for all $k \geq 0$, $x^{k+1} \leftarrow p_{H_k}(x^k)$ with sufficient decrease held at $p_{H_k}(x^k)$. Then the sequence $\{x^k\}$ satisfy

$$\Delta F_k := F(x^k) - F^* \leq \frac{2M^2(D_{\mathcal{X}_0} M + 2L_g)^2}{\rho\sigma^3} \frac{1}{k}.$$

Proof.

- $F(x^k) - F(x^{k+1}) \geq c\Delta F_k^2$,
- $\Delta F_k \leq c_1 \|\nabla f(x^k) + \gamma_g^{k+1}\|$,
- $F(x^k) - F(x^{k+1}) \geq c_2 \|\nabla f(x^k) + \gamma_g^{k+1}\|^2$.

Two Pillars!

Inexact Case

Theorem

Given $x^0 \in \mathbb{R}^n$, let the sequence $\{x^k\}$ be generated such that for all $k \geq 0$, $x^{k+1} \leftarrow p_{H_k}(x^k)$ with sufficient decrease held at $p_{H_k}(x^k)$. Then the sequence $\{x^k\}$ satisfy

$$\Delta F_k := F(x^k) - F^* \leq \frac{2M^2(D_{\mathcal{X}_0} M + 2L_g)^2}{\rho\sigma^3} \frac{1}{k}.$$

Proof.

- $F(x^k) - F(x^{k+1}) \geq c\Delta F_k^2$,
- $\Delta F_k \leq c_1 \|\nabla f(x^k) + \gamma_{g,\Phi}^{k+1}\|$,
- $F(x^k) - F(x^{k+1}) \geq c_2 \|\nabla f(x^k) + \gamma_{g,\Phi}^{k+1}\|^2$.

Inexact Case

Theorem

Given $x^0 \in \mathbb{R}^n$, let the sequence $\{x^k\}$ be generated such that for all $k \geq 0$, $x^{k+1} \leftarrow p_{H_k}(x^k)$ with sufficient decrease held at $p_{H_k}(x^k)$. Then the sequence $\{x^k\}$ satisfy

$$\Delta F_k := F(x^k) - F^* \leq \frac{2M^2(D_{\mathcal{X}_0} M + 2L_g)^2}{\rho\sigma^3} \frac{1}{k}.$$

Proof.

- $F(x^k) - F(x^{k+1}) \geq c\Delta F_k^2$,

- $\Delta F_k \leq c_1 \|\nabla f(x^k) + \gamma_{g,\Phi}^{k+1}\|$,
- $F(x^k) - F(x^{k+1}) \geq c_2 \|\nabla f(x^k) + \gamma_{g,\Phi}^{k+1}\|^2$.

NOT always true!

Two Pillars (Part 1)

Lemma

Consider $F(\cdot)$ *and any three points* $u, v, w \in \mathbf{dom}(F)$*, and we have*

$$F(u) - F(w) \leq \|\nabla f(u) + \gamma_{g,\phi}^{v}\| \|u - w\| + 2L_g \|u - v\| + 2\phi.$$

where $\gamma_{g,\phi}^{v} \in \partial_\phi g(v)$ *is any* ϕ*-subgradient of* $g(\cdot)$ *at point* v*.*

Remark

- consider $u = x^k, w = x^*$ and $v = x^{k+1}$
  - u – starting point
  - w – final point
  - v – point in the middle to connect u and w
- exact case $u == v$ implies optimality of $F(\cdot)!$ $(\phi = 0)$
  - with the first term $\|\nabla f(u) + \gamma_{g,\phi}^{v}\|$ also phased out, as we shall see later
- inexact case $u == v?$ $(\phi \neq 0)$

Two Pillars (Part 2)

Lemma

Let $x^{k+1} := p_{H_k,\phi_k}(x^k)$ *with some* $\phi_k \geq 0$. *Then*

$$Q(H_k, x^k, x^k) - Q(H_k, x^{k+1}, x^k) \geq \frac{\sigma_k}{2}\|x^{k+1} - x^k\|^2 - \sqrt{2M_k\phi_k}\|x^{k+1} - x^k\| - \phi_k.$$

Moreover there exists a vector $\gamma_{g,\phi}^{k+1} \in \partial g_{\phi_k}(x^{k+1})$ *such that the following bounds hold:*

$$\frac{1}{M_k}\|\nabla f(x^k) + \gamma_{g,\phi}^{k+1}\| - \frac{\sqrt{2M_k\phi_k}}{M_k} \leq \|x^{k+1} - x^k\| \leq \frac{1}{\sigma_k}\|\nabla f(x^k) + \gamma_{g,\phi}^{k+1}\| + \frac{\sqrt{2M_k\phi_k}}{\sigma_k}.$$

Remark

- especially useful when combined with sufficient decrease condition!
- inexact case the lower bound on $\|x^{k+1} - x^k\|$ might become trivial!

Inexact Case

Lemma

Consider kth iteration with $0 \leq \phi_k \leq 1$, $x^{k+1} := p_{H_k,\phi_k}(x^k)$ and $\Delta F_k := F(x^k) - F(x^*)$. Then there exists large enough positive constant $\theta > 0$, such that one of the following two cases must hold,

$$\Delta F_k \leq b_k \sqrt{\phi_k}, \tag{1.1}$$

$$\frac{1}{\Delta F_{k+1}} - \frac{1}{\Delta F_k} \geq c_k. \tag{1.2}$$

where b_k and c_k are given below,

$$b_k = \theta D_{\mathcal{X}_0} \sqrt{2M_k} + \frac{2(1+\theta)L_g}{\sigma_k} \sqrt{2M_k} + 2,$$

$$c_k = \frac{\rho(\sigma_k^3(\theta-1)^2 - 2\sigma_k M_k^2(1+\theta) - \sigma_k^3 M_k)}{(\sqrt{2}D_{\mathcal{X}_0}\theta\sigma_k M_k + 2\sqrt{2}L_g(1+\theta)M_k + \sigma_k\sqrt{M_k})^2}.$$

Inexact Case

Remark

- two cases corresponds to

$$\|\nabla f(x^k) + \gamma_{g,\phi}^{k+1}\| < \theta\sqrt{2M_k\phi_k} \Rightarrow (1.1),$$
$$\|\nabla f(x^k) + \gamma_{g,\phi}^{k+1}\| \geq \theta\sqrt{2M_k\phi_k} \Rightarrow (1.2).$$

- the lemma applies for any value of θ for which t_k, and hence, c_k is positive for all k.
- large θ imply large values of c_k, i.e., better rate w.r.t. (1.2).
- large θ is likely to cause both Case 1 to hold, i.e., (1.1), and a large b_k.
- the overall rate of convergence of the algorithm is derived using the two bounds - (1.1) and (1.2)
- the overall bound, thus, will depend on the upper bound on b_k's and the inverse of the lower bound on c_k's.
- If, again, we assume that $\sigma_k = M_k = L(f)$ for all k, then $\theta = O(\sqrt{L(f)})$ is sufficient to ensure that $c_k > 0$ and this results in $b_k \leq O(D_{\mathcal{X}_0} L(f))$ and $1/c_k \geq O(D^2_{\mathcal{X}_0} L(f))$, thus again, we obtain a bound which is comparable to that of proximal gradient methods, although with more complex constants.

Inexact Case

Lemma

Consider kth iteration with $0 \leq \phi_k \leq 1$, $x^{k+1} := p_{H_k,\phi_k}(x^k)$ and $\Delta F_k := F(x^k) - F(x^*)$. Then there exists large enough positive constant $\theta > 0$, such that one of the following two cases must hold,

$\|\nabla f(x^k) + \gamma_{g,\phi}^{k+1}\| < \theta\sqrt{2M_k\phi_k} \Rightarrow$ (1.1),

$\|\nabla f(x^k) + \gamma_{g,\phi}^{k+1}\| \geq \theta\sqrt{2M_k\phi_k} \Rightarrow$ (1.2).

$$\Delta F_k \leq b_k\sqrt{\phi_k}, \tag{1.1}$$

$$\frac{1}{\Delta F_{k+1}} - \frac{1}{\Delta F_k} \geq c_k. \tag{1.2}$$

where b_k and c_k are given below,

$$b_k = \theta D_{\mathcal{X}_0}\sqrt{2M_k} + \frac{2(1+\theta)L_g}{\sigma_k}\sqrt{2M_k} + 2,$$

$$c_k = \frac{\rho(\sigma_k^3(\theta-1)^2 - 2\sigma_k M_k^2(1+\theta) - \sigma_k^3 M_k)}{(\sqrt{2}D_{\mathcal{X}_0}\theta\sigma_k M_k + 2\sqrt{2}L_g(1+\theta)M_k + \sigma_k\sqrt{M_k})^2}.$$

Inexact Case

Lemma

Consider kth iteration with $0 \leq \phi_k \leq 1$, $x^{k+1} := p_{H_k,\phi_k}(x^k)$ and $\Delta F_k := F(x^k) - F(x^*)$. Then there exists large enough positive constant $\theta > 0$, such that one of the following two cases must hold,

same outer rate, but fewer inner steps!

$$\Delta F_k \leq b_k \sqrt{\phi_k}, \tag{1.1}$$

$$\frac{1}{\Delta F_{k+1}} - \frac{1}{\Delta F_k} \geq c_k. \tag{1.2}$$

where b_k and c_k are given below,

$$b_k = \theta D \chi_0 \sqrt{2M_k} + \frac{2(1+\theta)L_g}{\sigma_k}\sqrt{2M_k} + 2,$$

$$c_k = \frac{\rho(\sigma_k^3(\theta-1)^2 - 2\sigma_k M_k^2(1+\theta) - \sigma_k^3 M_k)}{(\sqrt{2}D\chi_0\theta\sigma_k M_k + 2\sqrt{2}L_g(1+\theta)M_k + \sigma_k\sqrt{M_k})^2}.$$

Inexact Case

Lemma

Consider kth iteration with $0 \leq \phi_k \leq 1, x^{k+1} := p_{H_k,\phi_k}(x^k)$ and $\Delta F_k := F(x^k) - F(x^*)$. Then there exists large enough positive constant $\theta > 0$, such that one of the following two cases must hold,

two horses tied running together!

$$\Delta F_k \leq b_k \sqrt{\phi_k}, \tag{1.1}$$

$$\frac{1}{\Delta F_{k+1}} - \frac{1}{\Delta F_k} \geq c_k. \tag{1.2}$$

where b_k and c_k are given below,

$$b_k = \theta D_{x_0} \sqrt{2M_k} + \frac{2(1+\theta)L_g}{\sigma_k} \sqrt{2M_k} + 2,$$

$$c_k = \frac{\rho(\sigma_k^3(\theta-1)^2 - 2\sigma_k M_k^2(1+\theta) - \sigma_k^3 M_k)}{(\sqrt{2}D_{x_0}\theta\sigma_k M_k + 2\sqrt{2}L_g(1+\theta)M_k + \sigma_k\sqrt{M_k})^2}.$$

Inexact Case

Theorem

Let the sequence $\{x^k\}$ *be generated such that for all* k, $x^{k+1} \leftarrow p_{H_k,\Phi_k}(x^k)$ *with sufficient decrease held at* $p_{H_k,\Phi_k}(x^k)$ *and with some* $\Phi_k \geq 0$ *that satisfy*

$$\phi_k \leq \frac{a^2}{k^2}, \text{ with } 0 < a \leq 1.$$

Let θ *be chosen as specified in the previous Lemma. Then for any* k

$$F(x^k) - F(x^*) \leq \frac{\max\{ba, \frac{1}{c}\}}{k-1}$$

faster than that there will be just constant improvement, and until a certain point!

Remark

- it follows that the inexact algorithm has sublinear convergence rate if $\phi_i \leq a^2/i^2$ for some $a < 1$ and all iterations $i = 0, \ldots, k$.
- in contrast, the analysis in [Schmidt et al., 2011] require that $\sum_{i=0}^{\infty} \sqrt{\phi_i}$ is bounded (and only applied to proximal gradient methods).
- this bound on the overall sequence is clearly stronger than $\phi_i \leq a^2/i^2$, since $\sum_{i=0}^{\infty} \frac{a}{i} = \infty$.
- on the other hand, it does not impose any particular requirement on any given iteration, except that each ϕ_i is finite, which our bound on ϕ_i is assumed to hold at each iteration, so far.

Inexact Case

Theorem

Let the sequence $\{x^k\}$ *be generated such that for all* k, $x^{k+1} \leftarrow p_{H_k,\Phi_k}(x^k)$ *with sufficient decrease held at* $p_{H_k,\Phi_k}(x^k)$ *and with some* $\phi_k \geq 0$ *that satisfy*

$$\phi_k \leq \frac{a^2}{k^2}, \text{ with } 0 < a \leq 1.$$

Let θ *be chosen as specified in the previous Lemma. Then for any* k

$$F(x^k) - F(x^*) \leq \frac{\max\{ba, \frac{1}{c}\}}{k-1}$$

Remark

- it follows that the inexact algorithm has sublinear convergence rate if $\phi_i \leq a^2/i^2$ for some $a < 1$ and all iterations $i = 0, \ldots, k$.
- in contrast, the analysis in [Schmidt et al., 2011] require that $\sum_{i=0}^{\infty} \sqrt{\phi_i}$ is bounded (and only applied to proximal gradient methods).
- this bound on the overall sequence is clearly stronger than $\phi_i \leq a^2/i^2$, since $\sum_{i=0}^{\infty} \frac{a}{i} = \infty$.
- on the other hand, it does not impose any particular requirement on any given iteration, except that each ϕ_i is finite, which our bound on ϕ_i is assumed to hold at each iteration, so far.

Total Complexity

Theorem

Suppose that at the k-th iteration function $Q(H_k, \cdot, x^k)$ is approximately minimized, to obtain x^{k+1} by applying $l(k) = \alpha k + \beta$ steps of any algorithm which guarantees that $Q(H_k, x^{k+1}, x^k) \leq Q(H_k, x^k, x^k)$ and whose convergence rate ensures the error bound $\phi_k \leq a^2/(\alpha k + \beta)^2$ for some $a > 0$. Then accuracy $F(x^k) - F(x^*) \leq \epsilon$ is achieved after at most

$$K = \beta\left(\frac{\max\{ba, \frac{1}{c}\}}{\epsilon} + 1\right) + \frac{\alpha}{2}\left(\frac{\max\{ba, \frac{1}{c}\}}{\epsilon}\right)\left(\frac{\max\{ba, \frac{1}{c}\}}{\epsilon} + 1\right)$$

inner iterations (of the chosen algorithm).

Remark

- $O(\frac{1}{\epsilon^2})$ inner steps!
- what about ϕ decreasing at linear rate? (recall Q is strongly convex!)

Total Complexity

Theorem

Suppose that at the k-th iteration function $Q(H_k, u, x^k)$ is approximately minimized, to obtain x^{k+1} by applying $l(k)$ steps of an algorithm, which guarantees that $Q(H_k, x^{k+1}, x^k) \leq Q(H_k, x^k, x^k)$ and whose convergence rate ensures the error bound $\phi_k \leq \delta^{l(k)} M_Q$, for some constants $0 < \delta < 1$ and $M_Q > 0$. Then, by setting $l_k = 2log_{\frac{1}{\delta}}(k)$, accuracy $F(x^k) - F(x^*) \leq \epsilon$ is achieved after at most

$$K = \sum_{k=0}^{t} 2 \log_{\frac{1}{\delta}}(k) \leq 2t \log_{\frac{1}{\delta}}(t)$$

inner iterations (of the chosen algorithm), with $t = \lceil \frac{\max\{ba, \frac{1}{c}\}}{\epsilon} + 1 \rceil$.

Remark

- $O(\frac{1}{\epsilon} \log(\frac{1}{\epsilon}))$ inner steps!
- $O(\frac{1}{\epsilon})$ outer steps for ISTA!
- inner problem is well-structured, i.e., quadratic + simple regularization
- outer problem, i.e., $F(\cdot)$, can be complicated!

Total Complexity

Theorem

Suppose that at the k-th iteration function $Q(H_k, u, x^k)$ is approximately minimized, to obtain x^{k+1} by applying $l(k)$ steps of an algorithm, which guarantees that $Q(H_k, x^{k+1}, x^k) \leq Q(H_k, x^k, x^k)$ and whose convergence rate ensures the error bound $\phi_k \leq \delta^{l(k)} M_Q$, for some constants $0 < \delta < 1$ and $M_Q > 0$. Then, by setting $l_k = 2log_{\frac{1}{\delta}}(k)$, accuracy $F(x^k) - F(x^*) \leq \epsilon$ is achieved after at most

$$K = \sum_{k=0}^{t} 2 \log_{\frac{1}{\delta}}(k) \leq 2t \log_{\frac{1}{\delta}}(t)$$

inner iterations (of the chosen algorithm), with $t = \lceil \frac{\max\{ba, \frac{1}{c}\}}{\epsilon} + 1 \rceil$.

Remark

- $O(\frac{1}{\epsilon} \log(\frac{1}{\epsilon}))$ inner steps!
- $O(\frac{1}{\epsilon})$ outer steps for ISTA!
- inner problem is well-structured, i.e., quadratic + simple regularization
- outer problem, i.e., $F(\cdot)$, can be complicated!
- each inner step exploits structure
- same steps as ISTA, but each step independent of n!
- locally super-linear rate!

Total Complexity

Theorem

Suppose that at the k-th iteration function $Q(H_k, u, x^k)$ is approximately minimized, to obtain x^{k+1} by applying $l(k)$ steps of an algorithm, which guarantees that $Q(H_k, x^{k+1}, x^k) \leq Q(H_k, x^k, x^k)$ and whose convergence rate ensures the error bound $\phi_k \leq \delta^{l(k)} M_Q$, for some constants $0 < \delta < 1$ and $M_Q > 0$. Then, by setting $l_k = 2log_{\frac{1}{\delta}}(k)$, accuracy $F(x^k) - F(x^*) \leq \epsilon$ is achieved after at most

$$K = \sum_{k=0}^{t} 2 \log_{\frac{1}{\delta}}(k) \leq 2t \log_{\frac{1}{\delta}}(t)$$

inner iterations (of the chosen algorithm), with $t = \lceil \frac{\max\{ba, \frac{1}{c}\}}{\epsilon} + 1 \rceil$.

Remark

- $O(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$ inner steps!
- $O(\frac{1}{\epsilon})$ outer steps for ISTA!
- inner problem is well-structured, i.e., quadratic + simple regularization
- outer problem, i.e., $F(\cdot)$, can be complicated!
- each inner step exploits structure
- same steps as ISTA, but each step independent of n!
- locally super-linear rate!

too good to be true!?

L-BFGS with low-rank structure

locally <u>super-linear</u> rate!

Randomized Coordinate Descent

each RCD step takes **constant time** <u>independent of data size!</u>

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…,l(k):
      Minimize local approximation
  Update variables
```

L-BFGS with low-rank structure

locally super-linear rate!

what is l(k)?
k?
log(k)?

Randomized Coordinate Descent

each RCD step takes **constant time** independent of data size!

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…,l(k):
      Minimize local approximation
  Update variables
```

L-BFGS with low-rank structure

locally super-linear rate!

what is l(k)?

n*log(k)!

Randomized Coordinate Descent

each RCD step takes **constant time** independent of data size!

```
For k = 1,2,…:
  Construct local approximation
  While direction is not good:
    Update local approximation
    For j = 1,2,…,l(k):
      Minimize local approximation
  Update variables
```

Probabilistic Case

Theorem

Let the sequence $\{x^k\}$ *be generated such that for all* k, $x^{k+1} \leftarrow p_{H_k,\phi_k}(x^k)$ *with sufficient decrease held at* $p_{H_k,\phi_k}(x^k)$ *and with some* $\phi_k \geq 0$ *that satisfy*

$$P\{\phi_k \leq \frac{a^2}{k^2}\} \geq 1-p, \text{ for some } 0 < a \leq 1 \text{ and } 0 \leq p < 1,$$

conditioned on the past. Let θ, b *and* c *be as specified in Theorem 12. Then for any* k

$$E(F(x^k) - F(x^*)) \leq \frac{\max\{ba, \frac{1}{c}\}(2-p)}{(1-p)(k-1)}.$$

Remark

- the expectation of ϕ_k needs to decrease at a rate faster than $O(\frac{1}{k^2})$.
- for $1-p$ percent of the time we have 'good' steps, with sufficient decrease on F.
- for the rest p percent of the time steps are 'bad'. But F still decreases.
- for large enough k, we will, eventually, have enough number of 'good' steps!

Probabilistic Case

Lemma

[Richtárik and Takáč, 2012] Let v be the initial point and $Q^* := \min_{u \in \mathbb{R}^n} Q(H, u, v)$. If v_l is the random point generated by applying l Randomized Coordinate Descent (RCD) steps to a strongly convex function Q, then for some constant we have

$$P\{Q(H, v_l, v) - Q^* \geq \phi\} \leq p,$$

as long as

$$i \geq n(1 + \mu(H)) \log(\frac{Q(H, v, v) - Q^*}{\phi p}),$$

where $\mu(H)$ is a constant that measures conditioning of H along the coordinate directions and in the worst case is at most M/σ - the condition number of H.

Remark

- RCD the expectation of ϕ_k decreases at a linear rate, i.e., $O(\delta^k)$.
- the constant $\delta = e^{-\frac{1}{n(1+\mu(H_k))}}$, which depends on n!
- hence, $l(k) = O(n(1 + \mu(H)) \log(kp/M_Q))$, which is $O(n \log(k))$!

Conclusions

- novel global analysis of Inexact Proximal Newton-like methods (IPN)
  - Schmidt et al. [2011] analyzes Inexact Proximal Gradient (IPG), which is a special case of IPN when H is diagonal, and which requires a stronger condition on error ϕ.
  - Jiang et al. [2012] also analyzes global rate for Proximal Newton, which requires much stricter conditions, i.e., $H_k - H_{k+1} \succeq 0$ (while providing FISTA-like rate).
  - Byrd et al. [2013] demonstrate super linear local convergence rate of the proximal Newton-like method (with the same sufficient decrease condition as ours, but applied within a line search).
- probabilistic analysis of RCD within IPN framework
- efficient algorithm combining RCD with L-BFGS within IPN framework
  - $O(\frac{n}{\epsilon}\log(\frac{1}{\epsilon}))$ RCD steps (yes, we do need n!)
  - the use of active-set can reduce n to nnz!
  - and each RCD step takes constant time!
  - and yes, we do notice super linear local convergence rate!
- provides theoretical guarantee to popular machine learning packages QUIC and LIBLINEAR.

Conclusions

- novel global analysis of Inexact Proximal Newton-like methods (IPN)
  - Schmidt et al. [2011] analyzes Inexact Proximal Gradient (IPG), which is a special case of IPN when H is diagonal, and which requires a stronger condition on error ϕ.
  - Jiang et al. [2012] also analyzes global rate for Proximal Newton, which requires much stricter conditions, i.e., $H_k - H_{k+1} \succeq 0$ (while providing FISTA-like rate).
  - Byrd et al. [2013] demonstrate super linear local convergence rate of the proximal Newton-like method (with the same sufficient decrease condition as ours, but applied within a line search).
- probabilistic analysis of RCD within IPN framework
- efficient algorithm combining RCD with L-BFGS within IPN framework
  - $O(\frac{n}{\epsilon}\log(\frac{1}{\epsilon}))$ RCD steps (yes, we do need n!)
  - the use of active-set can reduce n to nnz!
  - and each RCD step takes constant time!
  - and yes, we do notice super linear local convergence rate!
- provides theoretical guarantee to popular machine learning packages QUIC and LIBLINEAR.
- and a C/C++ implementation (with MATLAB and command line interface)
  - generic subproblem solver
  - generic objective interface
  - specialized L-BFGS compact representation library

Q&A

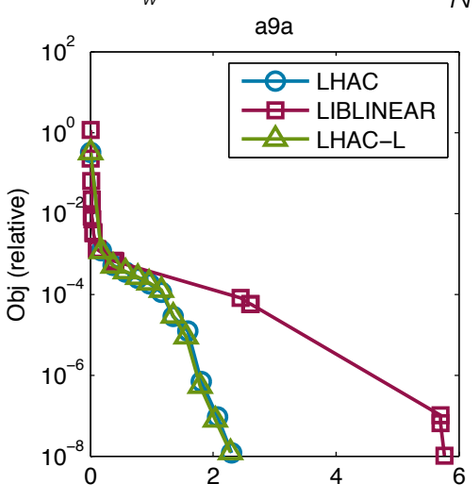
- can we apply L-BFGS to non-smooth function? (yes)
- can we combine L-BFGS with Randomized Coordinate Descent? (yes)
- is it faster than ISTA/FISTA? (yes)
- is it faster than L-BFGS with ISTA/FISTA? (yes)
- what is the complexity for each RCD step? (constant)
- how many RCD steps should we run per iteration? ($O(n \log(k))$)
- how many RCD steps do we need to achieve ϵ-accuracy? ($O(\frac{n}{\epsilon} \log(\frac{1}{\epsilon}))$)

LHAC

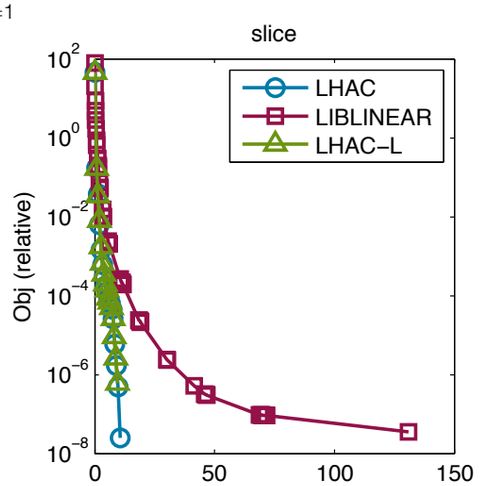
```
$ ./lhac.cmd -h
# output
Usage: lhac [options] training_set_file or model_file (see option m)
options:
-m model_file : model_file existence indicator (default false)
    true  -- read from model_file without training
    false -- train a new model from training_set_file
-p test file: apply model on the testing file
              and output the result to stdout
-d dense format : set matrix format dense or sparse (default 1)
    1  -- dense
    0  -- sparse
-l loss function : set type of loss function (default log)
    log    -- logistic regression
    square -- least square
-c lambda : set the regularization parameter (default 1)
-a : pre-compute A^TA in least sqaure  (default true)
-i : max number of iterations (default 1000)
-e epsilon : set tolerance of termination criterion
           final ista step size <= eps*(initial ista step size)
-v : set the verbose level (default 0)
    0 -- no output
    1 -- outer iteration
    2 -- sufficient decrease iteration
    3 -- coordinate descent iteration
```

LHAC

$$\min_{w} \quad F(w) = \lambda||w||_1 + \frac{1}{N}\sum_{n=1}^{N}\log(1+\exp(-y_n \cdot w^T x_n))$$

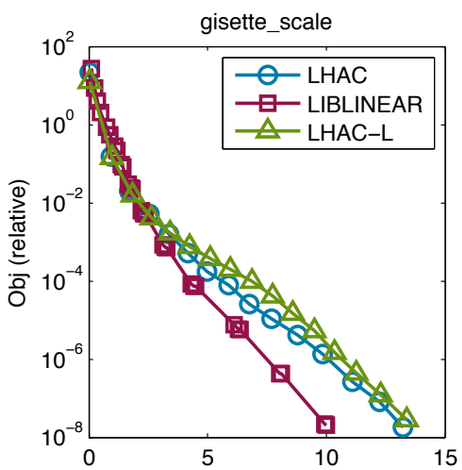


(c) a9a ($p = 123, N = 32561$)

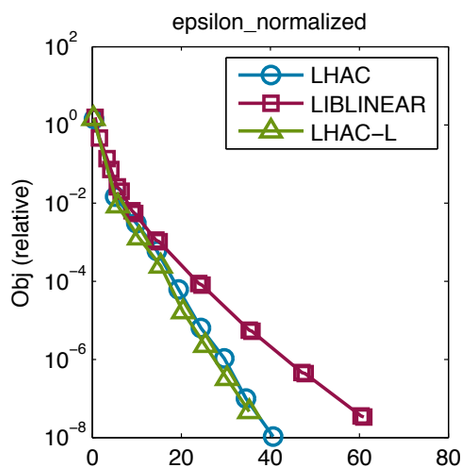


(d) slices ($p = 385, N = 53500$)

tol = 10^{-7} **(the y-axes on log scale).**



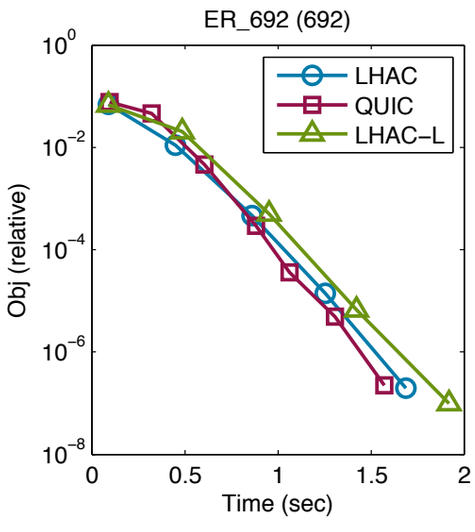
(a) gisette ($p = 5000, N = 6000$)



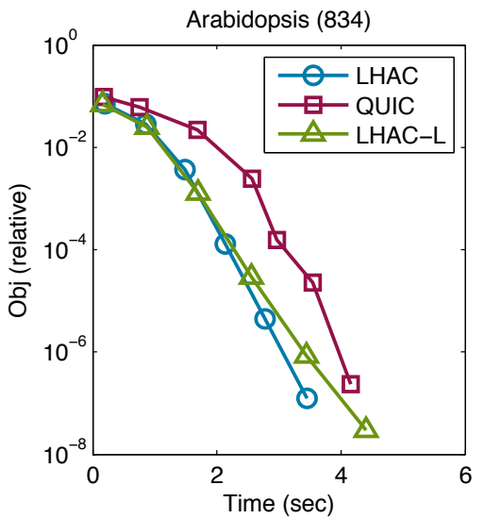
(b) epsilon ($p = 2000, N = 100000$)

tol = 10^{-7} **(the y-axes on log scale).**

$$\min_{X \succ 0} \quad F(X) = -\log\det X + \mathrm{tr}(SX) + \lambda||X||_1$$

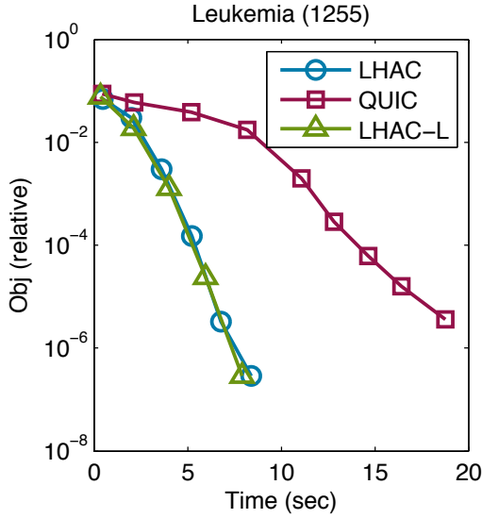


(a) $S : 692 \times 692$

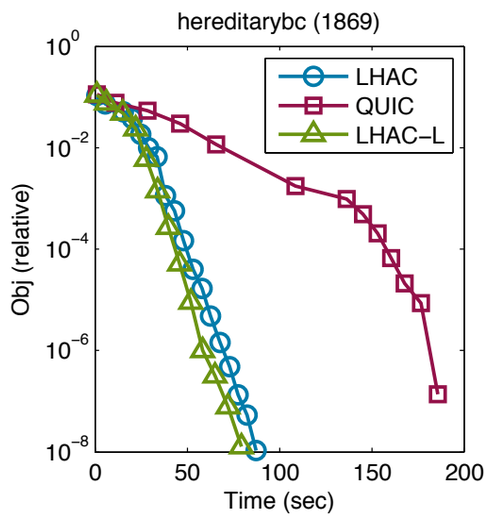


(b) $S : 834 \times 834$

tol = 10^{-7} **(the y-axes on log scale).**



(a) $S : 1255 \times 1255$



(b) $S : 1869 \times 1869$

tol = 10^{-7} **(the y-axes on log scale).**

Practical Inexact Proximal Quasi-Newton Method with Global Complexity Analysis

- arXiv:1311.6547
- detailed algorithm descriptions
- global convergence rate analysis under different scenarios
- experiment results

Thank you!

Questions?