Convex Optimization Algorithms for Machine Learning in 10 Slides

Presenter: Ian E.H. Yen

Jul. 15. 2015

Outline

1. Quadratic Problem—Linear System
2. Smooth Problem—Newton-CG
3. Composite Problem — Proximal-Newton-CD
4. Non-smooth, Non-separable—Augmented Lagrangian Method

Quadratic Problem

- Problem: $\min_w \ f(w) = \frac{1}{2} w^T H w + g^T w + c$
- Example:

$$\min_w \ f(w) = \frac{1}{2}\|y - Xw\|^2 + \frac{1}{2}\|w\|^2 \tag{1}$$

- Solution: solve a linear system

$$\nabla f(w) = 0 \Rightarrow Hw = -g \tag{2}$$

- How to solve?
  - In the example, $H = X^T X + I$ and $g = -Xy$.
  - X is $n \times d \rightarrow$ solving linear system directly requires $O(d^3)$.
  - **Hessian-vector product** (Hv) only needs $O(nnz(X))$.
- **Conjugate Gradient (CG)** produces reasonable solution using few iters of Hessian-vector product.

Smooth Problem — Newton-CG

- Problem: $\min_w f(w)$, where $\nabla f(w)$, $\nabla^2 f(w)$ are **continuous**.
- Ex.

$$\min_w f(w) = \sum_{i=1}^{n} L(w^T x_i, y_i) + \frac{1}{2}\|w\|^2 \tag{3}$$

  where $L(z, y) = \ln(1 + \exp(-yz))$ is logistic loss [1].
- **Newton-CG**, where each iter t we solve a quadratic approximation to find the "Newton direction" Δw_{nt}

$$\Delta w_{nt} = \underset{\Delta w}{argmin}\ \frac{1}{2}\Delta w^T H_t \Delta w + g_t^T \Delta w + f(w_t),$$

  and do line search to find step size η_t. $(w_{t+1} = w_t + \eta_t \Delta w_{nt})$
- In (4), $\mathbf{H_t = X^T D X + I}$, where D is diagonal matrix with $D_{ii} = L''(w_t^T x_i, y_i)$. $\Rightarrow$ Hessian-vector product $O(nnz(X))$.

[1] Lin. et al.. Trust region Newton method for logistic regression. ICML 2007.

Outline

1. Quadratic Problem—Linear System
2. Smooth Problem—Newton-CG
3. Composite Problem — Proximal-Newton-CD
4. Non-smooth, Non-separable—Augmented Lagrangian Method

Composite Problem

- Problem: $\min_w f(w) + h(w)$, where $f(w)$ is smooth, $h(w)$ is not smooth but **separable** w.r.t. "atoms".
- Ex. LASSO, L1-regularized Logistic Reg. [2]

$$\min_w f(w) = \sum_{i=1}^{n} L(w^T x_i, y_i) + \lambda \|w\|_1, \tag{4}$$

- Ex. Dual of SVM [3]

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q \alpha - \sum_{i=1}^{n} \alpha_i \\ s.t. \quad & 0 \leq \alpha_i \leq C, i = 1..n. \end{aligned} \tag{5}$$

- Ex. Matrix Completion [4]

$$\min_W \quad \frac{1}{2} \sum_{i,j \in \Omega} (A_{ij} - W_{ij})^2 + \lambda \|W\|_* \tag{6}$$

[2] Yuan et al. "An improved glmnet for l1-regularized logistic regression." JMLR 2012.

[3] Hsieh et al. "A dual coordinate descent method for large-scale linear SVM." ICML, 2008.

[4] Hsieh et al.. "Nuclear norm minimization via active subspace selection." ICML 2014.

Composite Problem

- Problem: $\min_w f(w) + h(w)$, where $f(w)$ is **smooth**, $h(w)$ is not smooth but **separable** w.r.t. "atoms".
- Insight: if $f(w)$ is **"atomic"** quadratic function, composite problem is easy to solve. Ex.

$$sign(\frac{-b}{a}) softThd(\frac{-b}{a}, \frac{\lambda}{a}) = \underset{x}{argmin} \ \frac{a}{2}x^2 + bx + \lambda|x|.$$

(Google "proximal operator for ..." to find formula you need.)

- **Proximal-Newton-CD:**

1. Construct local quadratic approximation $q(\Delta w; w_t)$. Solve

$$\Delta w^* = \underset{w}{argmin}\ q(\Delta w; w_t) + h(\Delta w + w_t). \tag{7}$$

via Coordinate Descent (optimize w.r.t. one atom at a time).

2. Do line search to find η_t and $w_{t+1} = w_t + \eta_t \Delta w^*$.

Composite Problem

Problem: $\min_w f(w) + h(w)$, where $f(w)$ is smooth, $h(w)$ is not smooth but **separable** w.r.t. "atoms".

- **Proximal-Newton-CD:**
  1. Construct local quadratic approximation $q(\Delta w; w_t)$. Solve

$$\Delta w^* = \underset{w}{argmin}\ q(\Delta w; w_t) + h(\Delta w + w_t). \tag{8}$$

  via Coordinate Descent (optimize w.r.t. one atom at a time).
  2. Do line search to find η_t and $w_{t+1} = w_t + \eta_t \Delta w^*$.
- **Key to efficiency:** whether $\nabla q(.) = H_t \Delta w + g_t$ can be maintained efficiently after coordinate update.
- What if not? (ex. Multiclass, CRF) $\Rightarrow$ **Prox-Quasi-Newton**: replace H_t with low-rank approximation B_t constructed from historical $\nabla f(w_1), ..., \nabla f(w_{t-1})$. [5]

[5] Zhong et al. "Proximal Quasi-Newton for Computationally Intensive l1-regularized M-estimators." NIPS 2014.

Outline

1. Quadratic Problem—Linear System
2. Smooth Problem—Newton-CG
3. Composite Problem — Proximal-Newton-CD
4. Non-smooth, Non-separable—Augmented Lagrangian Method

Non-smooth, Non-separable Problem

- What if the non-smooth function is non-separable ?
- Linear Program:

$$\begin{aligned} \min_{x,\xi \geq 0} \quad & c^T x \\ s.t. \quad & Ax + \xi = b. \end{aligned} \tag{9}$$

- Robust PCA:

$$\begin{aligned} \min_{L,S} \quad & \|L\|_* + \lambda \|S\|_1 \\ s.t. \quad & L + S = X \end{aligned} \tag{10}$$

- Reduce it to **composite problem** by Augmented Lagrangian Method !

Non-smooth, Non-separable Problem

- What if the non-smooth function is non-separable ?
- Linear Program:

$$\begin{aligned} \min_{x,\xi \geq 0} \quad & c^T x \\ s.t. \quad & Ax + \xi = b. \end{aligned} \tag{11}$$

- min-max of Lagrangian: (dual variable α)

$$\min_{x,\xi \geq 0} \max_{\alpha} \; c^T x + \alpha^T (Ax - b + \xi) \tag{12}$$

- **Augmented Lagrangian Method:**

$$\begin{aligned} (x^*, \xi^*) &= \underset{x,\xi \geq 0}{argmin} \; c^T x + \alpha_t^T (Ax - b + \xi) + \frac{1}{2} \|Ax - b + \xi\|^2 \\ \alpha_{t+1} &= \alpha_t + (Ax^* - b + \xi^*) \end{aligned} \tag{13}$$