Problem of interest

Consider the optimization problem

$$\underset{x \in \mathcal{X}}{\text{minimize}} \;\; F(x;\lambda) = f(x) + \lambda \|x\|_1 \tag{14}$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function
- $\lambda > 0$ is the regularization parameter

- we may replace $\|\cdot\|_1$ with a more general convex regularizer $r(x)$
- we could use the cutting plane, bundle, or trust-region method from the previous section because they do take advantage of convexity
- the cutting plane, bundle, and trust-region methods do not take advantage of the specific structure of the convex sparsifier $\|\cdot\|_1$
- we consider a gradient method that takes advantage of the "simplicity" of $\|\cdot\|_1$
- why do we care about problems of this form?
  - $f(x) = \|Ax - b\|_2^2$, where $A \in \mathbb{R}^{m \times n}$ with $m \ll n$, and we aim to find a solution x with lots of zeros
  - f may be a logistic-regression model whose parameters are obtained using maximum likelihood estimation, where the $\|\cdot\|_1$ gives preference to sparse predictors

Notes

A straightforward application of the subgradient method from the previous section gives the basic iteration

$$x_{k+1} = x_k - \tau_k \gamma_k g(x_k;\lambda) \tag{15}$$

where

$$g(x_k;\lambda) \in \nabla f(x_k) + \lambda \partial \|x_k\|_1$$

and

$$[\partial \|x\|_1]_i = \begin{cases} 1 & \text{if } x_i > 0 \\ -1 & \text{if } x_i < 0 \\ [-1,1] & \text{if } x_i = 0. \end{cases}$$

If we wanted a steepest subgradient descent method, we would use the minimum norm element of the subgradient, i.e.,

$$x_{k+1} \leftarrow x_k - \tau_k \gamma_k g_k^s \tag{16}$$

where

$$[g_k^s]_i = \begin{cases} [\nabla f(x_k)]_i + \lambda & \text{if } [x_k]_i > 0 \\ [\nabla f(x_k)]_i - \lambda & \text{if } [x_k]_i < 0 \\ S_\lambda([\nabla f(x_k)]_i) & \text{if } [x_k]_i = 0 \end{cases}$$

where the shrinkage operator S_λ is defined as

$$S_\lambda(y) = (|y| - \lambda)_+ \,\mathbf{sign}(y) = \begin{cases} y - \lambda & \text{if } y > \lambda \\ y + \lambda & \text{if } y < -\lambda \\ 0 & \text{if } |y| \le \lambda \end{cases}$$

Notes

We may observe that if we define x_{k+1} as the solution of

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x_k) + g(x_k;\lambda)^T(x - x_k) + \frac{1}{2\gamma_k\tau_k}\|x - x_k\|_2^2$$

then it satisfies

$$g(x_k;\lambda) + \frac{1}{\tau_k\gamma_k}(x_{k+1} - x_k) = 0$$

and after solving for x_{k+1} yields

$$x_{k+1} = x_k - \tau_k\gamma_k g(x_k;\lambda) \tag{17}$$

which is equivalent to (15).

- the previous iteration uses the structure of $\|\cdot\|_1$
- can we utilize the regularizer $\|\cdot\|_1$ even more since it is relatively "simple"?

Notes

Consider the subproblem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2\gamma_k\tau_k}\|x - x_k\|_2^2 + \lambda\|x\|_1 \tag{18}$$

- does not linearize the regularizer $\|\cdot\|_1$, but rather keeps it explicitly
- one might suspect that this subproblem will identify zeros more efficiently than (17)
- subproblem (18) might appear to be more difficult to solve, but it is not!

From the previous slide, the model to be minimized during the kth iteration is

$$m_k(x) = f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2\gamma_k\tau_k}\|x - x_k\|_2^2 + \lambda\|x\|_1 \tag{19}$$

and optimality conditions tell us that x_{k+1} satisfies

$$0 \in \partial m_k(x_{k+1}) = \nabla f(x_k) + \frac{1}{\tau_k\gamma_k}(x_{k+1} - x_k) + \lambda\partial\|x_{k+1}\|_1$$

which holds if and only if

$$0 = \nabla f(x_k) + \frac{1}{\tau_k\gamma_k}(x_{k+1} - x_k) + \lambda s_{k+1}$$

where

$$[s_{k+1}]_i \begin{cases} = -1 & \text{if } [x_{k+1}]_i < 0 \\ = 1 & \text{if } [x_{k+1}]_i > 0 \\ \in [-1, 1] & \text{if } [x_{k+1}]_i = 0 \end{cases}$$

and after rearrangement is equivalent to

$$x_{k+1} = x_k - \tau_k\gamma_k\nabla f(x_k) - \tau_k\gamma_k\lambda s_{k+1}.$$

Case 1: $[x_{k+1}]_i = 0$

$[x_{k+1}]_i = 0$ if and only if

$$[x_k - \tau_k\gamma_k\nabla f(x_k)]_i = \tau_k\gamma_k\lambda[s_{k+1}]_i \quad \text{for some } [s_{k+1}]_i \in [-1, 1]$$

which holds if and only if

$$[x_k - \tau_k\gamma_k\nabla f(x_k)]_i \in \tau_k\gamma_k\lambda[-1, 1].$$

Notes

Case 2: $[x_{k+1}]_i > 0$

$[x_{k+1}]_i > 0$ if and only if

$$0 < [x_{k+1}]_i = [x_k - \tau_k\gamma_k \nabla f(x_k)]_i - \tau_k\gamma_k\lambda$$

which holds if and only if

$$[x_k - \tau_k\gamma_k \nabla f(x_k)]_i > \tau_k\gamma_k\lambda.$$

Case 3: $[x_{k+1}]_i < 0$

$[x_{k+1}]_i < 0$ if and only if

$$0 > [x_{k+1}]_i = [x_k - \tau_k\gamma_k \nabla f(x_k)]_i + \tau_k\gamma_k\lambda$$

which holds if and only if

$$[x_k - \tau_k\gamma_k \nabla f(x_k)]_i < -\tau_k\gamma_k\lambda.$$

Summary

The minimizer x_{k+1} of m_k satisfies

$$[x_{k+1}]_i = \begin{cases} [x_k - \tau_k\gamma_k \nabla f(x_k)]_i - \tau_k\gamma_k\lambda & \text{if } [x_k - \tau_k\gamma_k \nabla f(x_k)]_i > \tau_k\gamma_k\lambda \\ [x_k - \tau_k\gamma_k \nabla f(x_k)]_i + \tau_k\gamma_k\lambda & \text{if } [x_k - \tau_k\gamma_k \nabla f(x_k)]_i < -\tau_k\gamma_k\lambda \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

which is equivalent to

$$x_{k+1} = S_{\tau_k\gamma_k\lambda}\big(x_k - \tau_k\gamma_k \nabla f(x_k)\big). \tag{21}$$

Notes

Iteration (21) is the basis for ISTA (Iterative Shrinkage-Thresholding Algorithm).

Question: How does the basic ISTA iteration (21) relate to the steepest descent iteration (16)?

Answer: If x_{k+1}^{SD} denotes the update associated with (16) and x_{k+1}^{ISTA} denotes the update associated with (20), then it may be shown that

- if $[x_k]_i > 0$ then

$$[x_{k+1}^{ISTA}]_i = \begin{cases} [x_{k+1}^{SD}]_i & \text{if } [x_{k+1}^{SD}]_i > 0 \\ [x_{k+1}^{SD}]_i + 2\tau_k\gamma_k\lambda & \text{if } [x_{k+1}^{SD}]_i < -2\tau_k\gamma_k\lambda \\ 0 & \text{otherwise} \end{cases}$$

- if $[x_k]_i < 0$ then

$$[x_{k+1}^{ISTA}]_i = \begin{cases} [x_{k+1}^{SD}]_i & \text{if } [x_{k+1}^{SD}]_i < 0 \\ [x_{k+1}^{SD}]_i - 2\tau_k\gamma_k\lambda & \text{if } [x_{k+1}^{SD}]_i > 2\tau_k\gamma_k\lambda \\ 0 & \text{otherwise} \end{cases}$$

- if $[x_k]_i = 0$ then

$$[x_{k+1}^{ISTA}]_i = [x_{k+1}^{SD}]_i$$

Notes

Instead of the bundle method master subproblem

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \; m_k(x) + \frac{\rho}{2}\|x - w_k\|_2^2$$

we use the trust-region master subproblem

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \; m_k(x) \quad \text{subject to} \quad \|x - w_k\| \le \delta$$

- $m_k(x) = \max_{j \in \mathcal{J}_k} \{\ell_j(x)\}$
- $\mathcal{J}_k$ is the kth index set
- $\ell_j(x) = f(x_j) + g_j^T(x - x_j)$
- $\delta > 0$ is the trust-region radius
- w_k is the best point so far
- the trust-region constraint restricts the search for a better point to a neighborhood of the previous best point w_k
- most common norm is $\|\cdot\|_\infty$
  - $\|x\|_\infty \le \delta \iff -\delta \le x_i \le \delta$ for all $1 \le i \le n$
  - if $\mathcal{X}$ is polyhedral, then the trust-region subproblem is equivalent to a smooth quadratic program, i.e., quadratic objective and linear constraints
  - if $\mathcal{X}$ is polyhedral and $\|\cdot\| = \|\cdot\|_2$, then the master subproblem would have a quadratic objective, linear constraints, and a single quadratic constraint
- we will use $\|\cdot\| = \|\cdot\|_\infty$

Notes

Rather than solving the nonsmooth master subproblem

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \; m_k(x) \quad \text{subject to} \quad \|x - w_k\|_\infty \le \delta$$

we solve the equivalent smooth master subproblem

$$\begin{aligned}(x_{k+1}, v_{k+1}) = \operatorname*{argmin}_{x \in \mathcal{X}, v \in \mathbb{R}} \; & v \\ \text{subject to} \quad & \ell_j(x) \le v \quad \text{for } j \in \mathcal{J}_k \\ & [w_k]_i - \delta \le x_i \le [w_k]_i + \delta \quad \text{for } 1 \le i \le n\end{aligned} \tag{13}$$

Notes