Lecture 22

First-order methods for structured problems

May 7, 2015

Massachusetts Institute of Technology
Lecture Outline

- Motivation
  - Structured problems: smooth + nonsmooth
  - Applications: regularized least-squares, etc.

- Algorithms
  - Proximal gradient methods
  - Accelerated gradient
Structured problems

- Convex optimization problems of the form

\[ \min_x f(x) + g(x) \]

where \( f(x) \) is smooth, but \( g(x) \) is nonsmooth.

- Many applications, often (very) large-scale

- Convex problem, can apply known methods (e.g., gradient, interior-point, etc)

- But, there’s additional structure that we could (should!) exploit...
Examples

• $\ell_1$-regularized least-squares

$$\min_x \|Ax - b\|^2 + \lambda \|x\|_1 \quad \|x\|_1 := \sum_i |x_i|$$

  - Least-squares problem, with an $\ell_1$ penalty
  - Other possibilities: wavelets, total variation norm (image processing), etc...

• Matrix versions, e.g.,

$$\min_x \|A(X)\|^2 + \lambda \|X\|_* \quad \|X\|_* := \sum_i \sigma_i(X)$$

  - $A : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p$ is an affine map
  - $\| \cdot \|_*$ is the nuclear norm (sum of singular values)
  - E.g., matrix completion (“Netflix”) problem
Special cases

Convex optimization, with $f(x)$ smooth and $g(x)$ nonsmooth:

$$\min_x f(x) + g(x)$$

For suitable choices of $f$ and $g$, we obtain:

- Smooth convex optimization ($g(x) = 0$)
- “Pure” nonsmooth convex optimization ($f(x) = 0$)
- Constrained smooth convex optimization ($g(x) = I_C(x)$, where $C$ is a convex set and $I_C$ is its indicator function).
- “Standard” constraints $g_i(x) \leq 0$ (convex). Pick $g(x) = \sum_i I_{(-\infty,0]}(g_i(x))$. 
How to solve this?

Let’s reinterpret gradient descent method (with fixed stepsize) for minimization of a smooth $f(x)$.

Consider a quadratic (or “regularized linear”) approximation

$$F_t(x; y) := f(y) + \nabla f(y)^T(x - y) + \frac{1}{2t} \|x - y\|^2$$

for some fixed $t > 0$. Notice this is strictly convex.

- Minimizing this approximation (over $x$) yields the minimizer

$$x^* = y - t\nabla f(y)$$

- This is the gradient method, with constant stepsize!

Interesting...
Gradient projection

What if we have a constraint $x \in C$, where $C$ is a convex set?

- Can rewrite (completion of squares) as:

$$F_t(x; y) := f(y) + \nabla f(y)^T (x - y) + \frac{1}{2t} \| x - y \|^2$$

$$= \frac{1}{2t} \| x - (y - t\nabla f(y)) \|^2 + f(y) - \frac{t}{2} \| \nabla f(y) \|^2$$

- Minimizing over $x \in C$ yields

$$x^* = [y - t\nabla f(y)]^+$$

where as before, $[\cdot]^+$ is the projection on the convex set $C$.

Gradient projection method!
Back to $f(x) + g(x)$

- Let's try the same approach here, leaving nonsmooth term as is...

$$F_t(x; y) := f(y) + \nabla f(y)^T (x - y) + \frac{1}{2t} \|x - y\|^2 + g(x)$$

and minimizing over $x$.

- Can rewrite as

$$F_t(x; y) = g(x) + \frac{1}{2t} \|x - (y - t\nabla f(y))\|^2 + w(y)$$

- Want to minimize over $x$
The proximal operator

This suggests to define the *prox operator* $\text{prox}_t(g) : \mathbb{R}^n \rightarrow \mathbb{R}^n$:

$$\text{prox}_t(g)(z) := \arg\min_u g(u) + \frac{1}{2t} \|u - z\|^2$$

- A convex optimization problem, optimality condition:
  $$0 \in \partial g(u^*) + \frac{1}{t}(u^* - z) \iff z \in u^* + t\partial g(u^*)$$

- Recall we have
  $$F_t(x; y) = g(x) + \frac{1}{2t} \|x - (y - t\nabla f(y))\|^2 + w(y)$$

- Using this, we can write its minimizer as
  $$x^* = \text{prox}_t(g)(y - t\nabla f(y))$$

  “take a step along the gradient of $f$, then apply the prox of $g$”.

This gives the *proximal gradient method*
Proximal gradient method

\[ x_{k+1} = \text{prox}_t(g)(x_k - t\nabla f(x_k)) \]

with the prox operator given by:

\[ \text{prox}_t(g)(z) := \arg\min_u g(u) + \frac{1}{2t}||u - z||^2 \]

Let’s check that this is sensible...

- If \( g(x) = 0 \), then \( \text{prox}_t(g)(z) = z \) (thus, “standard” gradient)
- If \( g(x) = I_C(x) \), then \( \text{prox}_t(g)(z) = [z]^+ \) (thus, gradient projection)
- What if \( g(x) \) is smooth and strictly convex?
Nonsmooth examples

Only makes sense whenever we can *efficiently* compute the prox:

\[
\text{prox}_t(g)(z) := \arg\min_u g(u) + \frac{1}{2t}\|u - z\|^2
\]

Example: \(\ell_1\) norm

- The \(\ell_1\) norm is \(g(x) = \sum_i |x_i|\). For the prox, need to solve

\[
\min_u \sum_i |u_i| + \frac{1}{2t}(u_i - z_i)^2.
\]

- Since this is separable, we obtain a “shrinkage” operator:

\[
[\text{prox}_t(g)(z)]_i = \phi_t(z_i), \quad \phi_t(z) := \begin{cases} 
z + t & \text{if } z \leq -t \\
0 & \text{if } |z| \leq t \\
z - t & \text{if } z \geq t. \end{cases}
\]
Prox map for nuclear norm

If \( g(X) = \|X\|_* \), then from the matrix SVD decomposition:

\[
Z = U \Sigma V^T, \quad \text{prox}_t(g)(Z) = U \phi_t(\Sigma) V^T,
\]

where \( \phi_t(\Sigma) = \text{diag}(\phi_t(\sigma_i)) \), i.e., “shrinkage” of singular values.

- Can use to solve, e.g., the matrix completion problem
Analysis of prox method

Similar to other cases we have seen. For simplicity, assume:

- The gradient of $f(x)$ is Lipschitz, i.e.,

  \[ \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \quad \forall x, y \]

- Constant stepsize $t = 1/L$, so method is:

  \[ x_{k+1} = \text{prox}_{1/L}(g) \left[ x_k - \frac{1}{L} \nabla f(x_k) \right] \]

- From Lipschitz property ("Descent lemma") obtain an upper bound:

  \[ f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \| y - x \|^2 \quad \forall x, y \]
• Let \( x_{k+1} = x_k - \frac{1}{L} \xi(x_k) \), i.e.,
\[
\xi(x) := L \left( x - \text{prox}_{1/L}(g) \left[ x - \frac{1}{L} \nabla f(x) \right] \right)
\]

• From Descent Lemma, an inequality for \( f \):
\[
f(x - \frac{1}{L} \xi(x)) \leq f(x) - \frac{1}{L} \nabla f(x)^T \xi(x) + \frac{1}{2L} \| \xi(x) \|^2
\]
where

• This yields an inequality for \( F = f + g \):
\[
F(x - \frac{1}{L} \xi(x)) \leq F(z) + \xi(x)^T (x - z) - \frac{1}{2L} \| \xi(x) \|^2 \quad \forall z
\]
(using convexity of \( f \) and of \( g \), and optimality condition
\( \xi(x) \in \nabla f(x) + \partial g(x - \frac{1}{L} \xi(x)) \))
• With $z = x$, this gives monotonicity:

$$F(x_{k+1}) \leq F(x_k) - \frac{1}{2L} \|\xi(x_k)\|^2$$

• With $z = x^*$, this is:

$$F(x_{k+1}) - F(x^*) \leq \xi(x_k)^T (x_k - x^*) - \frac{1}{2L} \|\xi(x_k)\|^2$$

$$= \frac{L}{2} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2)$$

In telescopic form!

Adding up, and using monotonicity, we’re done...
Convergence rate of prox method

For constant stepsize, can show a global bound:

\[ F(x_k) - F(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k} \]

- In this version, need (bound on) the Lipschitz constant.
- Sublinear rate, number of iterations is \( O(1/\epsilon) \).
Summary

- Splitting and prox methods nicely extend first-order methods to structured problems
- Need to be able to efficiently compute prox map
- Many variations exist (different stepsizes, line search, nonquadratic penalties, etc...)
- Can be “accelerated” using Nesterov’s fast gradient techniques.
- A few good references:
  - A. Beck, M. Teboulle, Gradient-based algorithms with applications to signal recovery problems.