Practical Session on Convex Optimization: Constrained Optimization

Mark Schmidt

INRIA/ENS

September 2011
Often we have constraints on problem:

- Natural bounds on the variables.
- Regularization or identifiability.
- Domain of function is restricted.

We may *introduce* constraints to use problem structure.
Example: $\ell_1$-Regularized Optimization

- $\ell_1$-regularization problems are of the form
  \[
  \min_w f(w) + \lambda \|w\|_1
  \]
  
- The problem is non-smooth because of the $\ell_1$-norm.
Example: $\ell_1$-Regularized Optimization

- $\ell_1$-regularization problems are of the form
  \[
  \min_w f(w) + \lambda \|w\|_1
  \]

- The problem is non-smooth because of the $\ell_1$-norm.
- We can convert this to a smooth constrained optimization:
  \[
  \min_{-s \leq w \leq s} f(w) + \lambda \sum_i s.
  \]
Example: $\ell_1$-Regularized Optimization

- $\ell_1$-regularization problems are of the form

$$\min_w f(w) + \lambda \|w\|_1$$

- The problem is non-smooth because of the $\ell_1$-norm.
- We can convert this to a smooth constrained optimization:

$$\min_{-s \leq w \leq s} f(w) + \lambda \sum_i s.$$ 

- Or write it as a smooth bound-constrained problem:

$$\min_{w^+ \geq 0, w^- \geq 0} f(w^+ - w^-) + \lambda \sum_i w^+ + \lambda \sum_i w^-.$$
Outline: Optimizing with Constraints

- Penalty-type methods for constrained optimization.
- Projection-type methods for constrained optimization.
- Convex Duality
Penalty-type methods re-write as an unconstrained problem, e.g.

- **Penalty method for equality constraints**: Re-write

\[
\min_{c(x)=0} f(x),
\]

as

\[
\min_x f(x) + \frac{\mu}{2} \|c(x)\|^2.
\]
Penalty-type methods re-write as an unconstrained problem, e.g.

- **Penalty method for equality constraints**: Re-write

\[
\min_{c(x)=0} f(x),
\]

as

\[
\min_x f(x) + \frac{\mu}{2} \|c(x)\|^2.
\]

- **Penalty method for inequality constraints**: Re-write

\[
\min_{c(x)\geq 0} f(x),
\]

as

\[
\min_x f(x) + \frac{\mu}{2} \|\max\{0, c(x)\}\|^2.
\]
Penalty-type methods re-write as an unconstrained problem, e.g.

- **Penalty method for equality constraints**: Re-write

  \[ \min_{c(x)=0} f(x), \]
  \[ \text{as} \]
  \[ \min_x f(x) + \frac{\mu}{2} \| c(x) \|^2. \]

- **Penalty method for inequality constraints**: Re-write

  \[ \min_{c(x)\geq 0} f(x), \]
  \[ \text{as} \]
  \[ \min_x f(x) + \frac{\mu}{2} \| \max\{0, c(x)\} \|^2. \]

  These converge to the original problem as \( \mu \to \infty \).
Penalty method for non-negative $\ell_1$-regularized logistic regression:

\[
\min_w \lambda \|w\|_1 + \sum_{i=1}^{n} \log(1 + \exp(-y_i(w^T x_i))).
\]

- Use an existing unconstrained optimization code.
- Solve for an increasing sequence of $\mu$ values.
Penalty method for non-negative $\ell_1$-regularized logistic regression:

$$\min_w \lambda ||w||_1 + \sum_{i=1}^{n} \log(1 + \exp(-y_i(w^T x_i))).$$

- Use an existing unconstrained optimization code.
- Solve for an increasing sequence of $\mu$ values.

Note the trade-off associated with penalty methods:
- Small $\mu$: easily solved but is a poor approximation.
- Large $\mu$ good approximation and is hard to solve.
Penalty method for non-negative $\ell_1$-regularized logistic regression:

$$\min_w \lambda \|w\|_1 + \sum_{i=1}^n \log(1 + \exp(-y_i(w^T x_i))).$$

- Use an existing unconstrained optimization code.
- Solve for an increasing sequence of $\mu$ values.

Note the trade-off associated with penalty methods:
- Small $\mu$: easily solved but is a poor approximation.
- Large $\mu$: good approximation and is hard to solve.

**Augmented Lagrangian** methods incorporate Lagrange multiplier estimates to improve the approximation for finite $\mu$. 

Mark Schmidt  
MLSS 2011  
Constrained Optimization
Augmented Lagrangian method for equality constraints:

1. Approximately solve

\[
\min_x f(x) + y_k^T c(x) + \frac{\mu}{2} \|c(x)\|^2.
\]

2. Update Lagrange multiplier estimates:

\[
y_{k+1} = y_k + \mu c(x).
\]

(for increasing sequence of \(\mu\) values)
**Augmented Lagrangian** method for inequality constraints:

1. Approximately solve:

   \[
   \min_x f(x) + y_k^T c(x) + \frac{\mu}{2} \left\| \max\{0, c(x)\} \right\|^2.
   \]

2. Update Lagrange multiplier estimates:

   \[
   y_{k+1} = \max\{0, y_k + \mu c(x)\}.
   \]

   (for increasing sequence of \(\mu\) values)
**Augmented Lagrangian Method**

**Augmented Lagrangian** method for inequality constraints:

1. Approximately solve

\[
\min_x f(x) + y_k^T c(x) + \frac{\mu}{2} \| \max\{0, c(x)\} \|^2.
\]

2. Update Lagrange multiplier estimates:

\[
y_{k+1} = \max\{0, y_k + \mu c(x)\}.
\]

(for increasing sequence of \(\mu\) values)

Exercise: Extend the penalty method to an augmented Lagrangian method.

Mark Schmidt MLSS 2011

Constrained Optimization
Exact penalty methods use a non-smooth penalty,

\[ \min_x f(x) + \mu \|c(x)\|_1, \]

and are equivalent to the original problem for finite \( \mu \).

Log-Barrier methods enforce strict feasibility,

\[ \min_x f(x) + \mu \sum_i \log c_i(x). \]

Most interior-point software packages implement a primal-dual log-barrier method.
Penalty-type methods for constrained optimization.
Projection-type methods for constrained optimization.
Convex Duality
Projection-type methods address the problem of optimizing over convex sets.

- A convex set $\mathcal{C}$ is a set such that

$$\theta x + (1 - \theta) y \in \mathcal{C},$$

for all $x, y \in \mathcal{C}$ and $0 \leq \theta \leq 1$. 
Projection-type methods address the problem of optimizing over convex sets.

- A convex set $C$ is a set such that
  \[ \theta x + (1 - \theta)y \in C, \]
  for all $x, y \in C$ and $0 \leq \theta \leq 1$.
- Projection-type methods use the projection operator,
  \[ P_C(x) = \arg \min_{y \in C} \frac{1}{2} \| x - y \|^2. \]
Projection-type Methods

Projection-type methods address the problem of optimizing over convex sets.

- A convex set $C$ is a set such that
  \[ \theta x + (1 - \theta)y \in C, \]
  for all $x, y \in C$ and $0 \leq \theta \leq 1$.

- Projection-type methods use the projection operator,
  \[ P_C(x) = \arg \min_{y \in C} \frac{1}{2}||x - y||^2. \]

- For non-negative constraints, this operator is simply
  \[ x = \max\{0, x\}. \]
The most basic projection-type method is *gradient projection*:

\[ x_{k+1} = P_C(x_k - \alpha_k \nabla f(x_k)). \]
The most basic projection-type method is gradient projection:

\[ x_{k+1} = P_C(x_k - \alpha_k \nabla f(x_k)). \]

We can use a variant of the Armijo condition to choose \( \alpha_k \):

\[ f(x_{k+1}) \leq f(x_k) - \gamma \nabla f(x_k)^T (x_{k+1} - x_k). \]

This algorithm has similar convergence and rate of convergence properties to the gradient method.

We can use many of the same tricks (polynomial interpolation, Nesterov extrapolation, Barzilai-Borwein step length).
The most basic projection-type method is gradient projection:

\[ x_{k+1} = P_C(x_k - \alpha_k \nabla f(x_k)) \].

We can use a variant of the Armijo condition to choose \( \alpha_k \):

\[ f(x_{k+1}) \leq f(x_k) - \gamma \nabla f(x_k)^T (x_{k+1} - x_k) \].

This algorithm has similar convergence and rate of convergence properties to the gradient method.

We can use many of the same tricks (polynomial interpolation, Nesterov extrapolation, Barzilai-Borwein step length).

Modify either the Nesterov code or the gradient code from the first session to do gradient projection for non-negative \( \ell_1 \)-regularized logistic regression.
There also exist projected-Newton methods where

\[ x_{k+1} = \arg\min_y \nabla f(x_k)^T (y - x_k) + \frac{1}{2\alpha_k} (y - x_k)^T \nabla^2 f(x_k) (y - x_k), \]

and analogous quasi-Newton and Hessian-free Newton methods.

Unfortunately, this problem is usually hard to solve.
There also exist projected-Newton methods where

\[ x_{k+1} = \arg \min_y \nabla f(x_k)^T (y-x_k) + \frac{1}{2\alpha_k} (y-x_k)^T \nabla^2 f(x_k) (y-x_k), \]

and analogous quasi-Newton and Hessian-free Newton methods.

Unfortunately, this problem is usually hard to solve.

But several heuristics are available:

1. **Sequential quadratic programming**: Use a linear approximation to the constraints.
2. **Active-Set**: Sub-optimize over a manifold of selected constraints.
3. **Two-metric projection**: Use a diagonal or other structured approximation to \( \nabla^2 f(x_k) \).
4. **Inexact projected-Newton**: Approximately compute \( x_{k+1} \).
Outline: Optimizing with Constraints

- Penalty-type methods for constrained optimization.
- Projection-type methods for constrained optimization.
- Convex Duality.
For the equality-constrained problem
\[
\min_{c(x)} f(x),
\]
the Lagrangian is defined as
\[
L(x, y) = f(x) + y^T c(x).
\]
For the equality-constrained problem

$$\min_{c(x)} f(x),$$

the Lagrangian is defined as

$$L(x, y) = f(x) + y^T c(x).$$

The Lagrange dual function is defined as

$$g(y) = \inf_x L(x, y),$$

and its domain is all values for which the infimum is finite.
The maximum of the dual lower bounds the primal,

\[ g(y^*) \leq f(x^*). \]
The maximum of the dual lower bounds the primal,

$$g(y^*) \leq f(x^*).$$

If \( g(y^*) = f(x^*) \), we say that strong duality holds.

Slater’s condition: for convex problems strong duality holds if a strictly feasible point exists.
Exercise: equality constrained norm minimization

Derive the dual function for the least-norm problem

\[
\min_{x} x^T x. \\
\text{subject to } Ax = b
\]
Derive the dual function for the least-norm problem

\[ \min_{x} x^T x \]

subject to \( Ax = b \)

1. Write out the Lagrange dual function.
2. Solve for \( x \).
3. Plug in the solution.
The **convex conjugate** \( f^* \) of function \( f \) is defined as

\[
f^*(y) = \sup_x (y^T x - f(x)),
\]

and its domain is all values for which the supremum is finite.

**Examples:**

1. If \( f(x) = \frac{1}{2} x^T x \), then \( f^*(y) = \frac{1}{2} y^T y \).
The convex conjugate $f^*$ of function $f$ is defined as

$$f^*(y) = \sup_x (y^T x - f(x)),$$

and its domain is all values for which the supremum is finite.

Examples:

1. If $f(x) = \frac{1}{2} x^T x$, then $f^*(y) = \frac{1}{2} y^T y$.
2. If $f(x) = ax + b$, then $f^*(y) = -b$ and $y = a$. 
The convex conjugate $f^*$ of function $f$ is defined as

$$f^*(y) = \sup_x (y^T x - f(x)),$$

and its domain is all values for which the supremum is finite.

Examples:

1. If $f(x) = \frac{1}{2} x^T x$, then $f^*(y) = \frac{1}{2} y^T y$.
2. If $f(x) = ax + b$, then $f^*(y) = -b$ and $y = a$.
3. If $f(x) = \log \sum_i e^{x_i}$, then $f^*(y) = \sum_i y_i \log y_i$ for $y \geq 0$ and $\sum_i y_i = 1$. 
Conjugate functions

- The convex conjugate $f^*$ of function $f$ is defined as

$$f^*(y) = \sup_x (y^T x - f(x)),$$

and its domain is all values for which the supremum is finite.

Examples:

1. If $f(x) = \frac{1}{2} x^T x$, then $f^*(y) = \frac{1}{2} y^T y$.
2. If $f(x) = ax + b$, then $f^*(y) = -b$ and $y = a$.
3. If $f(x) = \log \sum_i e^{x_i}$, then $f^*(y) = \sum_i y_i \log y_i$ for $y \geq 0$ and $\sum_i y_i = 1$.
4. If $f(x) = \|x\|_p$, then $f^*(y) = 0$ for $\|y\|_q \leq 1$. 
Exercise: equality constrained norm minimization 2

Derive the dual function for the least-norm problem

$$\min_{Ax=b} \|x\|_p.$$
In some cases we *introduce* constraints to derive a dual.

Derive a dual of the $\ell_1$-regularization problem

$$\min_x ||Ax - b||^2 + ||x||_1,$$

by re-formulating as

$$\min_{x, r = Ax - b} ||r||^2 + ||x||_1.$$
Similarly, the graphical LASSO problem

$$\min_X \log \det X + tr(X\Sigma) + \lambda\|X\|_1,$$

for $X$ positive-definite can be re-written as

$$\min_{\lambda \leq Y \leq \lambda} \log \det Y,$$

for $Y$ positive-definite.

Modify the projected-gradient code to solve the graphical LASSO problem (ignore the positive-definite constraint).
Most of this lecture is based on material from Nocedal and Wright’s very good “Numerical Optimization” book, and from Boyd and Vandenberghe’s very good “Convex Optimization” book.