Examples of convex programs

Sivaraman Balakrishnan
Outline

- Introduction to convex programming
- Example 1 – Minimum volume ellipsoid
- Example 2 – Distance metric learning
- Example 3 – Graphical Lasso
- Efficient algorithms for the Graphical Lasso
Convex programming in a slide

\[
\begin{align*}
\min_{x} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0 \\
& \quad h_j(x) = 0
\end{align*}
\]

- f and g are convex
- h is affine
Special cases

- LP - \( f, g, h \) linear
- QP - \( f \) is quadratic, \( g, h \) linear
- SOCP - \( f \) is linear, \( g \) is of the form \[ \|Ax + b\| \leq c^T x + d \]
- SDP – where \( F \)s are symmetric matrices
  \[
  \min_x \quad c^T x \\
  \text{subject to} \quad F_0 + x_1 F_1 + x_2 F_2 + \ldots + x_n F_n \geq 0
  \]
Minimum volume ellipsoid

- Introduction
  - Ellipsoid algorithm

- Formulation
  - Given $m$ points in $\mathbb{R}^n$
  - Want to solve

\[
\begin{align*}
\min_{A, x_C} \quad & \text{vol}(A) \\
\text{subject to} \quad & (x_i - x_C)^T A (x_i - x_C) \leq 1 \quad \forall \ i \in \{1, 2, \ldots, m\} \\
& A \in S_n^+
\end{align*}
\]
Minimum volume ellipsoid

- Use fact \( \text{vol}(A) \propto \frac{1}{\sqrt{\det(A)}} \)

- Any thoughts on what the proportionality constant is?

- Formulation

\[
\max_{A, x_C} \log(\det(A)) \\
\text{subject to } (x_i - x_C)^T A (x_i - x_C) \leq 1 \hspace{1mm} \forall \hspace{1mm} i \in \{1, 2, \ldots, m\} \\
A \in S_n^+
\]
Distance metric learning

- Introduction
  - Clustering with side information

- Distance metric – metric defines distance between elements of a set (typically in $\mathbb{R}^n$)

- Example – Euclidean distance

- Generalization – Mahalanobis distance

- Idea is to learn distance metric and then run K-means or some other clustering algorithm

- Constrained K-means algorithm
Distance metric learning contd.

- **Formulation as a convex program**
  - First attempt
    \[
    \min_A \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2
    \]
    subject to \(A \geq 0\)
  - Actual convex program
    \[
    \min_A \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2
    \]
    subject to
    \[
    \sum_{(x_i, x_j) \in D} \|x_i - x_j\|_A \geq 1
    \]
    \(A \geq 0\)
Glasso preliminaries

- Conditional independence
- Sparse structure
  - Statistical efficiency
  - Interpretability
Glasso

- Formulation from class

$$\max_X P(Y_{1:n}|X) = \max_X \frac{n}{2} \log|X| - \frac{n}{2} p \log(2\pi) - \frac{\sum_i y_i^T X y_i}{2}$$

- More generally

$$\max_X P(Y_{1:n}|X) = \max_X \log|X| - \frac{\sum_i (y_i - \mu)^T X (y_i - \mu)}{n}$$

- “Trace trick”
Glasso contd.

Formulation

- Given \( n \) samples \( (y^{(1)}, y^{(2)}, \ldots, y^{(n)}) \) drawn from \( N(\mu, \Sigma) \)
- Samples in \( p \) dimensions
- Find the precision matrix

\[
\Lambda = \arg \max_{X \succeq 0} \log \det X - \text{trace}(SX) - \lambda \|X\|_1
\]

With sample covariance

\[
S = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \mu)(y^{(i)} - \mu)^T
\]
Glasso contd.

- Consider case with no penalty – simple closed form solution 😊
- What if S is not invertible? 😞
**Dual norm**

- Definition from class

\[
\|x\|_* = \sup_{\|y\| \leq 1} x^T y
\]

- But, I learned

\[
\|x\|_p \text{ is dual to } \|x\|_q \text{ if } \frac{1}{p} + \frac{1}{q} = 1
\]

- Holder’s inequality

\[
x^T y \leq \|x\|_p \|y\|_q
\]
Dual norms and dual “glasso”

- Dual norm \[ \|X\|_1 = \max_{\|U\|_\infty \leq 1} \text{trace}(XU) \]

- Dual problem \[ \min_{\|U\|_\infty \leq \lambda} -\log(\det(S + U)) - p \]

- Strictly speaking need to have (S+U) invertible but in practice this is redundant (can be more formal about this but it’s not necessary – see paper if interested)
Efficient solution by block coordinate descent

- High level idea
  - Convergence and separability

- Dual problem again

\[
\max \quad \log(\det(W))
\]
\[
||W - S||_\infty \leq \lambda
\]

- Schur complement

\[
W = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}
\]

\[
\det(W) = \det(A)\det(C - B^T A^{-1} B)
\]
Block “glasso” problem

- W initialized to S+lambda (diagonals never change)
- At each step solve

\[ \begin{align*}
\min_y & \quad y^T A^{-1} y \\
\text{subject to} & \quad ||y - S_j||_{\infty} \leq \lambda
\end{align*} \]

- Update W

- This is a simple box QP with (p-1) variables and can be solved very very efficiently !!!!
Why is it called “glasso”? 

- Dual of this Box QP is a LASSO problem which can also be solved efficiently.

- The LASSO problem gives some intuition of what the algorithm does and relates it to another method called “neighborhood selection.”