Optimization Models

EECS 127 / EECS 227AT

Laurent El Ghaoui

EECS department
UC Berkeley

Fall 2018

LECTURE 11

Convex Quadratic Programs

> *We next consider the rule that the investor does (or should) consider expected return a desirable thing and variance of return an undesirable thing.*
>
> Harry Markowitz

Outline

Introduction

- A quadratic function in a vector of variables $x = [x_1\, x_2\, \cdots\, x_n]$ can be written generically as

$$f_0(x) = \frac{1}{2}x^\top Hx + c^\top x + d \quad \text{(a quadratic function)}$$

  where $d \in \mathbb{R}$, $c \in \mathbb{R}^n$, and $H \in \mathbb{R}^{n,n}$ is a symmetric matrix.

- A linear function is of course a special case of a quadratic function, obtained considering $H = 0$:

$$f_0(x) = c^\top x + d \quad \text{(a linear function)}.$$

- Linear and quadratic models treated in this lecture take the form

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to:} \quad & A_{\mathrm{eq}}x = b_{\mathrm{eq}} \\ & f_i(x) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

  where $f_0, \ldots, f_m$ are either quadratic or linear functions. We consider the case when these functions are *convex*, which happens if and only if the Hessian matrices of f_i, $i = 0, 1, \ldots, m$, are positive semi-definite.

Unconstrained minimization of quadratic functions

The linear case

- Consider first the linear case, $f_0(x) = c^\top x + d$:

$$p^* = \min_{x \in \mathbb{R}^n} \; c^\top x + d.$$

- It is an intuitive fact that $p^* = -\infty$ (i.e., the objective is unbounded below) whenever $c \neq 0$, and $p^* = d$, otherwise.
- Indeed, for $c \neq 0$ one may take $x = -\alpha c$, for any $\alpha > 0$ large at will, and drive f_0 to $-\infty$. Contrary, for $c = 0$ the function is actually constant and equal to d.
- We have therefore, for a linear function,

$$p^* = \begin{cases} d & \text{if } c = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Unconstrained minimization of quadratic functions

The quadratic case

$$p^* = \min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top H x + c^\top x + d.$$

- H has a negative eigenvalue. Then, f_0 is unbounded below, i.e., $p^* = -\infty$.
- All eigenvalues of H are nonnegative: $\lambda_i \geq 0$, $i = 1, \ldots, n$ (i.e., f_0 is convex). Then, the minima are characterized by the condition that the gradient of the function is zero, that is

  $$\nabla f_0(x) = Hx + c = 0.$$

  The minimizer should thus satisfy the system of linear equations $Hx = -c$.

  - If $c \notin \mathcal{R}(H)$, then there is no minimizer (f_0 is unbounded below).
  - If $c \in \mathcal{R}(H)$, then $p^* = -\frac{1}{2} c^\top H^\dagger c + d$, attained at any $x^* = -H^\dagger c + \zeta$, $\zeta \in \mathcal{N}(H)$.
  - All eigenvalues of H are positive: $\lambda_i > 0$, $i = 1, \ldots, n$. Then, there exist a unique minimizer $x^* = -H^{-1} c$, with $p^* = -\frac{1}{2} c^\top H^{-1} c + d$.

Example: Least Squares

- We have already encountered a special case of quadratic minimization problem, in the context of the Least-Squares approximate solution of linear equations.
- Indeed, the LS problem amounts to minimizing $f_0(x) = \|Ax - y\|_2^2$, hence

$$f_0(x) = (Ax - y)^\top (Ax - y) = x^\top A^\top Ax - 2y^\top Ax + y^\top y,$$

  which is a quadratic function in the standard form (4), with

$$H = 2(A^\top A), \quad c = -2A^\top y, \quad d = y^\top y.$$

- Note that f_0 is always convex, since $A^\top A \succeq 0$.
- The solution is given by the first-order optimality condition. Since $c \in \mathcal{R}(H)$, a LS solution satisfying these conditions always exists.
- Further, if A is full rank, then $A^\top A \succ 0$, then the solution is unique and it is given by the well-known formula

$$x^* = -H^{-1}c = (A^\top A)^{-1}A^\top y.$$

Example: Quadratic minimization under linear equality constraints

- The linear equality-constrained problem

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to:} \quad & Ax = b, \end{aligned}$$

  with $f_0(x) = \frac{1}{2}x^\top Hx + c^\top x + d$, can be readily converted into unconstrained form by *eliminating* the equality constraints.

- Parameterize all x such that $Ax = b$ as $x = \bar{x} + Nz$, where $\bar{x}$ is one specific solution of $Ax = b$, N is a matrix containing by columns a basis for the nullspace of A, and z is a vector of free variables.

- Then, we substitute x in f_0 and obtain a problem which is unconstrained in the variable z:

$$\min_z \varphi_0(z) = \frac{1}{2}z^\top \bar{H}z + \bar{c}^\top z + \bar{d},$$

  where

$$\bar{H} = N^\top HN,\ \bar{c} = N^\top(c + H\bar{x}),\ \bar{d} = d + c^\top \bar{x} + \frac{1}{2}\bar{x}^\top H\bar{x}.$$

Quadratic programs

- A quadratic optimization problem (or quadratic program, QP) is one of standard form where f_0 is a quadratic function (4) and $f_1, \ldots, f_m$ are affine functions.
- The feasible set of QP is a polyhedron (as in LP), but the objective is quadratic, rather than linear.
- If the H matrix in (4) is positive semi-definite, then we have a convex QP.
- The standard form of a QP is thus

$$\begin{aligned} p^* = \min_x \quad & \tfrac{1}{2} x^\top H x + c^\top x \\ \text{s.t.:} \quad & A_{\mathrm{eq}} x = b_{\mathrm{eq}} \\ & Ax \le b. \end{aligned}$$

Quadratic programs

Example: Tracking a financial index

- Consider a financial portfolio design problem, where the entries of $x \in \mathbb{R}^n$ represent the fractions of an investor's total wealth invested in each of n different assets, and where $r(k) \in \mathbb{R}^n$ represents the vector of simple returns of the component assets during the k-th period of time $[(k-1)\Delta, k\Delta]$, where Δ is a fixed duration, e.g., one month.
- Suppose that the components y_k of vector $y \in \mathbb{R}^T$ represents the return of some target financial index over the k-th period, for $k = 1, \ldots, T$.
- the so-called *index tracking* problem is to construct a portfolio x so to track as close as possible the "benchmark" index returns y.
- Since the vector of portfolio returns over the considered time horizon is

$$z = Rx, \quad R \doteq \begin{bmatrix} r^\top(1) \\ \vdots \\ r^\top(T) \end{bmatrix} \in \mathbb{R}^{T,n}$$

we may seek for the portfolio x with minimum LS tracking error, by minimizing $\|Rx - y\|_2^2$.

Quadratic programs

Example: Tracking a financial index

- Elements of x represent relative *weights*, that is they are nonnegative and they sum up to one. The index tracking problem is therefore a constrained LS problem, thus a convex QP:

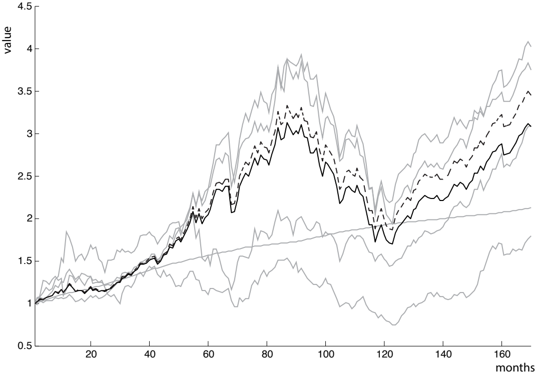
$$\begin{aligned} p^* = \min_x \quad & \|Rx - y\|_2^2 \\ \text{s.t.:} \quad & \mathbf{1}^\top x = 1, \quad x \geq 0. \end{aligned}$$

- 169 monthly return data of six indices: the MSCI US index, the MSCI EUR index, the MSCI JAP index, the MSCI PACIFIC index, the MSCI BOT liquidity index, and the MSCI WORLD index.
- The problem is to track the target index MSCI WORLD, using a portfolio composed by the other five indices.
- Solving the convex QP with this data, we obtain the optimal portfolio composition $x^* = [0.5138 \;\; 0.3077 \;\; 0.0985 \;\; 0.0374 \;\; 0.0426]^\top$, and hence the optimal-tracking portfolio return sequence $z^* = Rx^*$, with tracking error $\|Rx^* - y\|_2^2 = 2.6102 \times 10^{-4}$.

Quadratic programs

Example: Tracking a financial index

The figure below shows the result of investing one Euro into each of the component indices and benchmark index (solid line), and into the tracking-optimal portfolio. As expected, the value sequence generated by the optimal portfolio (dashed) is the closest one to the target index.



Quadratic constrained quadratic programs (QCQP)

- A generalization of the QP model is obtained by allowing quadratic (rather than merely linear) equality and inequality constraints. A quadratic constrained quadratic program (QCQP) thus takes the form

$$\begin{aligned} p^* = \min_x \quad & x^\top H_0 x + 2c_0^\top x + d_0 \\ \text{s.t.:} \quad & x^\top H_i x + 2c_i^\top x + d_i \leq 0 \quad i \in \mathcal{I} \\ & x^\top H_j x + 2c_j^\top x + d_j = 0 \quad j \in \mathcal{E}, \end{aligned} \tag{1}$$

  where $\mathcal{I}$, $\mathcal{E}$ denote the index sets relative to inequality constraints and equality constraints, respectively.
- A QCQP is convex if and only if $H_0 \succeq 0$, $H_i \succeq 0$ for $i \in \mathcal{I}$, and $H_j = 0$ for $j \in \mathcal{E}$. In other words, a QCQP is convex whenever the functions describing the objective and the inequality constraints are convex quadratic, and all the equality constraints are actually affine.

Problems involving cardinality and their ℓ_1 relaxations

- Many engineering applications require the determination of solutions that are *sparse*, that is possess only few nonzero entries (low-cardinality solutions).
- The quest for low-cardinality solutions often has a natural justification in terms of the general principle of *parsimony* of the ensuing design.
- However, finding minimum cardinality solutions (i.e., solutions with small ℓ_0 norm) is hard in general, from a computational point of view.
- For this reason, several *heuristics* are often used in order to devise tractable numerical schemes that provide low (albeit possibly not minimal) cardinality solutions. One of these schemes involves replacing the ℓ_0 norm with the ℓ_1 norm.

Problems involving cardinality and their ℓ_1 relaxations

- An interesting relation between the ℓ_1 norm of $x \in \mathbb{R}^n$ and its cardinality is obtained via the Cauchy-Schwartz inequality applied to the inner product of $|x|$ and $\mathrm{nz}\,(x)$, where $|x|$ is the vector whose entries are the absolute values of x, and $\mathrm{nz}\,(x)$ is the vector whose i-th entry is one whenever $x_i \neq 0$, and its is zero otherwise.
- For all $x \in \mathbb{R}^n$,

$$\|x\|_1 = \mathrm{nz}\,(x)^\top |x| \leq \|\mathrm{nz}\,(x)\|_2 \cdot \|x\|_2 = \|x\|_2 \sqrt{\mathrm{card}\,(x)},$$

  hence

$$\mathrm{card}\,(x) \leq k \quad \Rightarrow \quad \|x\|_1^2 \leq k\|x\|_2^2.$$

- Also, for all $x \in \mathbb{R}^n$,

$$\|x\|_1 = \mathrm{nz}\,(x)^\top |x| \leq \mathrm{nz}\,(x)^\top \mathbf{1} \cdot \|x\|_\infty = \|x\|_\infty \mathrm{card}\,(x),$$

  hence

$$\mathrm{card}\,(x) \leq k \quad \Rightarrow \quad \|x\|_1 \leq k\|x\|_\infty.$$

Piece-wise constant fitting

- Suppose one observes a noisy time-series which is almost piece-wise constant. The goal in piece-wise constant fitting is to find what the constant levels are. In biological or medical applications, such levels might have interpretations of "states" of the system under observation.
- Let $x \in \mathbb{R}^n$ denote the signal vector (which is unknown) and let $y \in \mathbb{R}^n$ denote the vector of noisy signal observations (i.e., y is true signal x, plus noise).
- Given y, we seek an estimate $\hat{x}$ of the original signal x, such that $\hat{x}$ has as few changes in consecutive time steps as possible.
- We model the latter requirement by minimizing the cardinality of the difference vector $D\hat{x}$, where $D \in \mathbb{R}^{n-1,n}$ is the difference matrix

$$D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix},$$

so that $D\hat{x} = [\hat{x}_2 - \hat{x}_1,\, \hat{x}_3 - \hat{x}_2,\, \ldots,\, \hat{x}_n - \hat{x}_{n-1}]^\top$.

Piece-wise constant fitting

- We are thus led to the problem

$$\min_{\hat{x}} \|y - \hat{x}\|_2^2 \quad \text{s.t.: } \operatorname{card}(D\hat{x}) \le k,$$

  where k is an estimate on the number of jumps in the signal.
- Here, the objective function in the problem is a measure of the error between the noisy measurement and its estimate $\hat{x}$.
- We can relax this hard problem via the ℓ_1-norm heuristic, by replacing the cardinality constraint with an ℓ_1 constraint, thus obtaining the QP

$$\min_{\hat{x}} \|y - \hat{x}\|_2^2 \quad \text{s.t.: } \|D\hat{x}\|_1 \le q,$$

  for some suitably chosen q (note that choosing $q = k$ need not imply that the solution resulting from the relaxed problem is such that $\operatorname{card}(D\hat{x}) \le k$).
- Alternatively, one may cast a problem with a weighted objective:

$$\min_{\hat{x}} \|y - \hat{x}\|_2^2 + \gamma \|D\hat{x}\|_1,$$

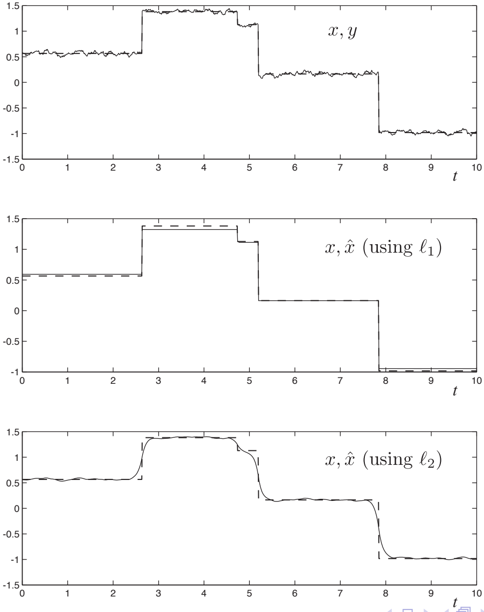
  for some suitable trade-off parameter $\gamma \ge 0$.

Piece-wise constant fitting

Example of signal reconstruction via piece-wise fitting.

- The top panel shows the unknown signal x (dashed) and its available noisy measurement y; the center panel shows the unknown signal x (dashed) and its reconstruction
- $\hat{x}$ obtained via the ℓ_1 heuristic; the bottom panel shows the unknown signal x (dashed) and its reconstruction $\hat{x}$ obtained by solving a regularization problem where the ℓ_2 norm is used instead of the ℓ_1 norm in the constraint.
- We notice that the ℓ_1 heuristic is successful in eliminating the noise from the signal, while preserving sharp transitions in the phase (level) changes in the signal.
- Contrary, with an ℓ_2 heuristic, noise elimination only comes at the price of sluggish phase transitions.

Piece-wise constant fitting



ℓ_1 regularization and the LASSO

- Regularized LS problems, with an ℓ_2 regularization term, have been discussed in Lecture 5.
- An important variation arises when the regularization term involves the ℓ_1 norm of x, instead of the ℓ_2 norm. This results in the following problem, known as the *basis pursuit denoising problem* (BPDN), or as the *least absolute shrinkage and selection operator* (LASSO) problem:

$$\min_{x \in \mathbb{R}^n} \|Ax - y\|_2^2 + \lambda \|x\|_1, \quad \lambda \geq 0, \tag{2}$$

  where $\|x\|_1 = |x_1| + \cdots + |x_n|$.
- Problem (2) received enormous attention in recent years from the scientific community, due to its relevance in the field of *compressed sensing* (CS).
- The basic idea is that the ℓ_1 norm of x is used as a proxy for the cardinality of x (the number of nonzero entries in x).
- It formalizes a tradeoff between the accuracy with which Ax approximates y, and the *complexity* of the solution, intended as the number of nonzero entries in x. The larger λ is, the more problem (2) is biased towards finding low-complexity solutions, i.e., solutions with many zeros.

ℓ_1 regularization and the LASSO

- Problem (2) can be cast in the form of a standard QP by introducing slack variables $u \in \mathbb{R}^n$:

$$\min_{x,u \in \mathbb{R}^n} \quad \|Ax - y\|_2^2 + \lambda \sum_{i=1}^{n} u_i$$
$$\text{s.t.:} \quad |x_i| \leq u_i, \; i = 1, \ldots, n.$$

- Typical applications where LASSO-type problems arise may involve a very large number of variables, hence several specialized algorithms have been developed to solve ℓ_1-regularized problems with maximal efficiency.

Image compression in a wavelet basis

- A gray-scale image, represented by a vector $y \in \mathbb{R}^m$, typically admits an essentially sparse representation, in a suitable basis:
- This means that, for appropriate dictionary matrix $A \in \mathbb{R}^{m,n}$, the image y can be well approximated by a linear combination Ax of the feature vectors, where the coefficients x of the combination are sparse.
- Usual dictionary matrices employed in image analysis include Discrete Fourier Transform (DFT) bases, and wavelet (WT) bases. Wavelet bases, in particular, have been recognized to be quite effective in providing sparse representations of standard images (they are used, for instance in the Jpeg2000 compression protocol).
- Consider, for example, the 256×256 gray-scale image shown next. Each pixel in this image is represented by an integer value y_i in the range [0, 255], where the 0 level is for black, and 255 is for white.

Image compression in a wavelet basis

- Original image, and histogram of y: non sparse!

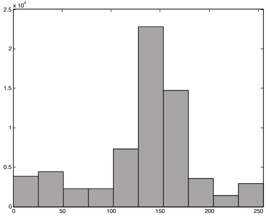


Image compression in a wavelet basis

- However, if we consider the image representation in the wavelet transform domain (which implicitly amounts to considering a suitable dictionary matrix A containing by columns the wavelet bases), we obtain a vector representation $\tilde{y}$ whose absolute value has the following histogram. For this example, we are using a Daubechies orthogonal wavelet transform, hence A is a 65536×65536 orthogonal matrix.

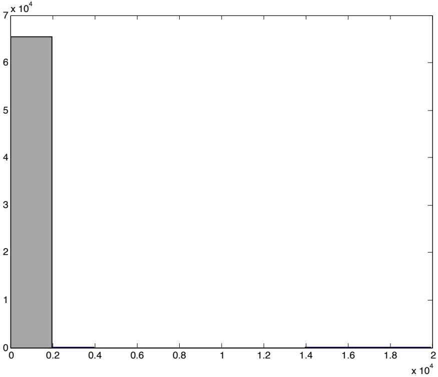


Image compression in a wavelet basis

- The wavelet representation $\tilde{y}$ of the image contains very few large coefficients, while most of the coefficient are relatively small (however, $\tilde{y}$ is not yet sparse, since its elements are not exactly zero).
- If all these small coefficients are retained, then $\tilde{y}$ carries the same information as y, that is, it is a *lossless* encoding of the original image, in the wavelet domain: $y = A\tilde{y}$.
- However, if we allow for this equality to be relaxed to approximate equality $y \simeq Ax$, we may tradeoff some accuracy in change of a representation x in the wavelet domain which has many zero coefficients, i.e., a sparse representation.
- Such a sparse tradeoff can typically be obtained by solving the LASSO problem (2) for suitable λ, that is $\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1$.

Image compression in a wavelet basis

- In our specific situation, since A is orthogonal, we have that the above problem is equivalent to
$$\min_x \; \frac{1}{2}\|x - \tilde{y}\|_2^2 + \lambda \|x\|_1,$$
where $\tilde{y} \doteq A^\top y$ is the image representation in the wavelet domain.
- This problem is *separable*, i.e., it can be reduced to a series of univariate minimization problems, since
$$\frac{1}{2}\|x - \tilde{y}\|_2^2 + \lambda \|x\|_1 = \sum_{i=1}^{m} \frac{1}{2}(x_i - \tilde{y}_i)^2 + \lambda |x_i|.$$
- Moreover, each of the single-variable problems
$$\min_{x_i} \; \frac{1}{2}(x_i - \tilde{y}_i)^2 + \lambda |x_i|$$
admits a simple closed-form solution as
$$x_i^* = \begin{cases} 0 & \text{if } |\tilde{y}_i| \leq \lambda \\ \tilde{y}_i - \lambda \mathrm{sgn}(\tilde{y}_i) & \text{otherwise.} \end{cases}$$

Image compression in a wavelet basis

- All coefficients $\tilde{y}_i$ in the wavelet basis are thresholded to zero if their modulus is smaller than λ, and are offset by λ, otherwise (soft thresholding).
- Once we computed x^*, we can reconstruct an actual image in the standard domain, by computing the inverse wavelet transform (i.e., ideally, we construct the product Ax^*).
- Solving the LASSO problem with $\lambda = 30$ we obtained a representation x^* in the wavelet domain that has only 4540 nonzero coefficients (against the 65536 nonzero coefficients present in $\tilde{y}$ or in y). We have therefore a compression factor of about 7%, meaning that the size of the compressed image is only 7% of the size of the original image.
- Reducing the regularization parameter to $\lambda = 10$, we obtained instead a representation x^* in the wavelet domain with 11431 nonzero coefficients, and thus a compression factor of about 17%

Image compression in a wavelet basis

Comparison of original boat image (a), wavelet compression with $\lambda = 10$ (b), and wavelet compression with $\lambda = 30$ (c).

(a)

(b)

(c)