ECON 7377 Microeconometrics **Spring 2026**

Optimization

Contents

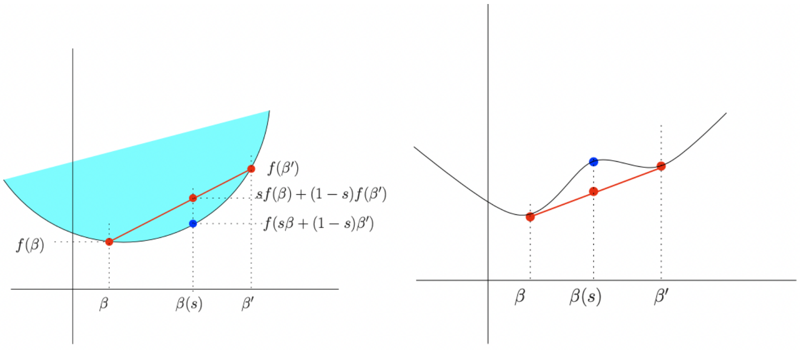
- We focus on convex optimization problems with convex objective functions and convex constraints,

$$\min_{\beta} f(\beta) \text{ s.t. } \beta \in \mathcal{C}$$

where $f : \mathbb{R}^p \mapsto \mathbb{R}$ is a convex function and $\mathcal{C} \subset \mathbb{R}^p$ is a convex constraint set.

  - Appendix A gives a brief summary on the characterization and properties of convex sets and functions. Boyd and Vandenberghe (2004) offers a comprehensive treatment.

- Any Local minimum is also a global minimum.



(a) (b)

Figure 5.1 *(a) For a convex function, the line $sf(\beta) + (1-s)f(\beta)$ always lies above the function value $f(s\beta + (1-s)\beta)$. (b) A nonconvex function that violates the inequality* (5.1). *Without convexity, there may be local minima that are not globally minima, as shown by the point β'.*

Proof. Suppose $\beta_1 \in \mathcal{C}$ is a local minimum, i.e. $\exists\, \delta > 0$ such that $\forall\, \beta \in B_\delta(\beta_1) \cap \mathcal{C}$, $f(\beta) \geq f(\beta_1)$; $\beta_0 \in \mathcal{C}$ is a global minimum with $f(\beta_1) > f(\beta_0)$. Construct $\bar{\beta} = \lambda\beta_0 + (1-\lambda)\beta_1$ with $\lambda = \frac{\delta}{2\|\beta_1 - \beta_0\|}$, then $\bar{\beta} \in B_\delta(\beta_1) \cap \mathcal{C}$ by convexity of $\mathcal{C}$. By convexity of f,

$$f\left(\bar{\beta}\right) \leq \lambda f(\beta_0) + (1-\lambda) f(\beta_1) < f(\beta_1),$$

which contradicts the assumption that β_1 is a local minimum. □

- A general convex optimization problem is often of the form

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^p} f(\beta) \\ \text{s.t. } & g_j(\beta) \leq 0, j = 1, 2, \cdots, m, \\ & A\beta - b = 0, \end{aligned} \tag{1}$$

where f, g_j, $j = 1, 2, \cdots, m$, are convex functions and $A \in \mathbb{R}^{q \times p}$.

  - The feasible set is characterized by sublevel sets of convex functions and affine sets.

1 Optimality Conditions

- To analyze and solve an optimization problem, we need to characterize the optimal solutions by a system of equations and inequalities (*optimality conditions*).

1.1 Unconstrained Optimization Problems

- A general unconstrained nonlinear optimization problem

$$\min_{\beta \in \mathbb{R}^p} f(\beta)$$

Proposition 1. Suppose that $f : \mathbb{R}^p \mapsto \mathbb{R}$ is **continuously differentiable** at $\bar{\beta} \in \mathbb{R}^p$, and $\exists\, d \in \mathbb{R}^p$ such that $\nabla f(\bar{\beta})^\top d < 0$, then $\exists\, \bar{\alpha} > 0$ such that $f\left(\bar{\beta} + \alpha d\right) < f\left(\bar{\beta}\right)$, $\forall\, \alpha \in (0, \bar{\alpha})$. In other words, d is a **descent direction** at $\bar{\beta}$.

Proof. $\exists\, \bar{\alpha} > 0$ such that $\nabla f\left(\bar{x} + ad\right)^\top d < 0$ for $\alpha \in (0, \bar{\alpha})$. $\forall\, \alpha \in (0, \bar{\alpha})$,

$$f\left(\bar{\beta} + \alpha d\right) = f\left(\bar{\beta}\right) + \alpha \nabla f\left(\bar{\beta} + td\right)^\top d < f\left(\bar{\beta}\right)$$

for some $t \in [0, \alpha)$. □

Corollary 1 (First order necessary condition for unconstrained optimization). *Suppose that $f : \mathbb{R}^p \mapsto \mathbb{R}$ is continuously differentiable at $\bar{\beta} \in \mathbb{R}^p$. If $\bar{\beta}$ is a local minimum, then $\nabla f\left(\bar{\beta}\right) = 0$. In particular, $\left\{d \in \mathbb{R}^p : \nabla f(\bar{\beta})^T d < 0\right\} = \emptyset$.*

Proof. Choose $d = -\nabla f\left(\bar{\beta}\right)$. □

- The necessary condition is also sufficient for convex functions.

Proposition 2. Suppose that $f : \mathbb{R}^p \mapsto \mathbb{R}$ is convex on S and continuously differentiable at $\bar{\beta} \in S$ where $S \subset \mathbb{R}^p$ is an open convex set, then $f\left(\bar{\beta}\right) \leq f(\beta)\ \forall\, \beta \in S$ if and only if $\nabla f\left(\bar{\beta}\right) = 0$.

Proof. Use the characterization $f(\beta) \geq f\left(\bar{\beta}\right) + \nabla f\left(\bar{\beta}\right)^\top \left(\beta - \bar{\beta}\right)$ of differentiable convex functions. □

- The key ingredients of the proof is that the convex function f is minorized by an affine function. The result is ready to be extended to non-differentiable functions.

Definition 1. For a convex function f, a vector z is a **subgradient** of f at $\bar{\beta}$ if $f(\beta) \geq f(\beta) + z^\top \left(\beta - \bar{\beta}\right)\ \forall\, \beta \in \mathbb{R}^p$. The set of all subgradients of f at $\bar{\beta}$ is call **subdifferential**, denoted by $\partial f\left(\bar{\beta}\right)$.

- The subgradient z defines a non-vertical supporting hyperplane of $\text{epi}\,(f)$ at $(\beta, f(\beta))$.

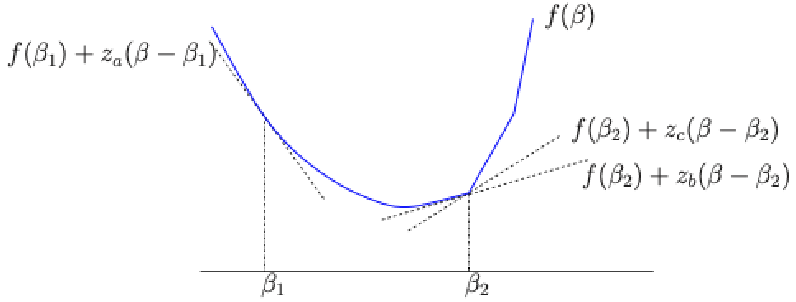


Figure 5.3 *A convex function $f : \mathbb{R} \to \mathbb{R}$, along with some examples of subgradients at β_1 and β_2.*

- Example 1: $\partial f(\beta) = \begin{cases} \{+1\} & \text{if } \beta > 0 \\ \{-1\} & \text{if } \beta < 0 \\ [-1, +1] & \text{if } \beta = 0 \end{cases}$ when $f(\beta) = |\beta|$.

- Example 2: $\partial \|\Theta\|_* = \left\{ Z \in \mathbb{R}^{m \times n} : Z = \sum_{j=1}^{\min\{m,n\}} z_j u_j v_j^\top, z_j \in \text{sign}\left(\sigma_j\left(\Theta\right)\right) \right\}$ where $\Theta \in \mathbb{R}^{m \times n}$ has the singular value decomposition $\Theta = \sum_{j=1}^{\min\{m,n\}} \sigma_j\left(\Theta\right) u_j v_j^\top$, and $\sigma_j\left(\Theta\right)$ is the j-th singular value of Θ.

Proposition 3. For any $f : \mathbb{R}^p \mapsto \mathbb{R}$, $f\left(\bar{\beta}\right) \leq f\left(\beta\right) \,\forall\, \beta \in \mathbb{R}^p$ if and only if $0 \in \partial f\left(\bar{\beta}\right)$.

- $\partial f\left(\beta\right)$ may not be easy to compute. For nonconvex differentiable f, we may not have $\nabla f\left(\beta\right) \in \partial f\left(\beta\right)$.

Proposition 4 (Second order sufficient condition for unconstrained optimization). Suppose that $f : \mathbb{R}^p \mapsto \mathbb{R}$ is twice continuously differentiable at $\bar{\beta} \in \mathbb{R}^p$. If $\nabla f\left(\bar{\beta}\right) = 0$ and $\nabla^2 f\left(\bar{x}\right) \succ 0$, then $\bar{\beta}$ is a local minimum.

1.2 Constrained optimization

- Consider a general optimization problem (not necessarily convex) with both equality and inequality constraints,

$$\begin{aligned} &\min_{\beta \in \mathcal{B}} f\left(\beta\right) \\ \text{s.t. } & g_j\left(\beta\right) \leq 0,\ j = 1, 2, \cdots, m_1, \\ & h_k\left(\beta\right) = 0,\ k = 1, 2, \cdots, m_2. \end{aligned} \tag{2}$$

where $\mathcal{B}$ is an open nonempty subset of $\mathbb{R}^p$. Denote the feasible set by

$$\mathcal{S} = \left\{\beta \in \mathcal{B} : g_j\left(\beta\right) \leq 0,\ j = 1, 2, \cdots, m_1, h_k\left(\beta\right) = 0,\ k = 1, 2, \cdots, m_2\right\}.$$

- Assume f, g_j, $j = 1, 2, \cdots, m_1$ and h_k, $k = 1, 2, \cdots, m_2$ are continuously differentiable on $\mathcal{S}$.

Theorem 2 **(Fritz John necessary condition).** *Let $\bar{\beta} \in \mathcal{S}$ be a local mininum of* (2)*. Then* $\exists$ $\rho \in \mathbb{R}$*,* $\lambda \in \mathbb{R}^{m_1}$*,* $\mu \in \mathbb{R}^{m_2}$ *such that*

$$\begin{aligned}
&\rho \nabla f\left(\bar{\beta}\right) + \sum_{j=1}^{m_1} \lambda_j \nabla g_j\left(\bar{\beta}\right) + \sum_{k=1}^{m_2} \mu_k \nabla h_k\left(\bar{\beta}\right) = 0, \\
&\rho, \lambda_j, \geq 0, j = 1, 2, \cdots, m_1, \\
&\left(\rho, \lambda^\top, \mu^\top\right)^\top \neq 0. \\
&\lambda_j g_j\left(\bar{\beta}\right) = 0,\ j = 1, 2, \cdots, m_1.
\end{aligned} \tag{3}$$

- Consider a special case where only one inequality constraint, $g\left(\beta\right) \leq 0$ is present. The Lagrangian condition becomes $-\nabla f\left(\bar{\beta}\right) = \frac{\lambda}{\rho} \nabla g\left(\bar{\beta}\right)$ if $\rho > 0$, for any $d \in \mathbb{R}^p$, $-\nabla f\left(\bar{\beta}\right)^\top d = \frac{\lambda}{\rho} \nabla g\left(\bar{\beta}\right)^\top d$, i.e. feasibility and descent cannot be achieved simultaneously.

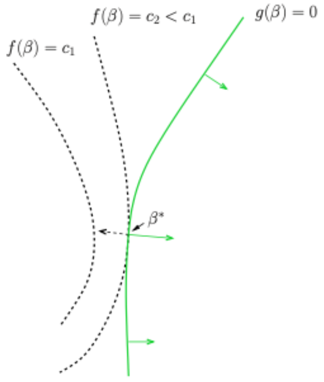


Figure 5.2 *Illustration of the method of Lagrange multipliers. We are minimizing a function f subject to a single constraint $g(\beta) \leq 0$. At an optimal solution β^*, the normal vector $\nabla f(\beta^*)$ to the level sets of the cost function f points in the opposite direction to the normal vector $\nabla g(\beta^*)$ of the constraint boundary $g(\beta) = 0$. Consequently, up to first order, the value of $f(\beta^*)$ cannot be decreased by moving along the contour $g(\beta) = 0$.*

Proof of Theorem 2. The proof adopts the penalty function approach (see Bertsekas (1999) for details). Define a sequence of penalized problems,

$$\min_{\beta \in B_\epsilon\left(\bar{\beta}\right)} F^q(\beta) \equiv f(\beta) + \frac{q}{2} \sum_{j=1}^{m_1} \left(g_j^+(\beta)\right)^2 + \frac{q}{2} \sum_{k=1}^{m_2} \left(h_k(\beta)\right)^2 + \frac{1}{2}\|\beta - \bar{\beta}\|_2^2, \tag{4}$$

where $g_j^+(x) = \max\{g_j(x), 0\}$, for $j = 1, 2, \cdots, m_1$, and $\epsilon > 0$ is chosen such that $f(\bar{\beta}) < f(\beta)$, $\forall \beta \in B_\epsilon(\bar{\beta}) \cap \mathcal{S}$. Let β^q be the optimal to (4) with q, where $q = 1, 2, \cdots$.

1. Show that $\beta^q \to \bar{\beta}$ as $q \to \infty$.

   Note that

   $$f(\beta^q) \leq F^q(\beta^q) = f(\beta^q) + \frac{q}{2}\sum_{j=1}^{m_1}\left(g_j^+(\beta^q)\right)^2 + \frac{q}{2}\sum_{k=1}^{m_2}\left(h_k(\beta^q)\right)^2 + \frac{1}{2}\|\beta^q - \bar{\beta}\|_2^2 \leq F^q(\bar{\beta}) = f(\bar{\beta}) \tag{5}$$

   where the last equality comes from feasibility of $\bar{\beta}$. Divide (5) by q, we have

   $$\frac{f(\beta^q)}{q} \leq \frac{F^q(\beta^q)}{q} = \frac{f(\beta^q)}{q} + \frac{1}{2}\sum_{j=1}^{m_1}\left(g_j^+(\beta^q)\right)^2 + \frac{1}{2}\sum_{k=1}^{m_2}\left(h_k(\beta^q)\right)^2 + \frac{1}{2q}\|\beta^q - \bar{\beta}\|_2^2 \leq \frac{f(\bar{\beta})}{q}.$$

   $f(\beta^q)$ is bounded over $B_\epsilon(\bar{\beta})$ and $\|\beta^q - \bar{\beta}\|_2^2 \leq \epsilon^2$, $q = 1, 2, \cdots$, then

   $$\lim_{q\to\infty}\sum_{j=1}^{m_1}\left(g_j^+(\beta^q)\right)^2 = \lim_{q\to\infty}\sum_{k=1}^{m_2}\left(h_k(\beta^q)\right)^2 = 0.$$

   Let $\tilde{\beta}$ be a limit point of $\{\beta^q\}$, then $\tilde{\beta} \in \mathcal{S} \cap B_\epsilon(\bar{\beta})$. By (5) and continuity of f ,

   $$f\left(\tilde{\beta}\right) + \frac{1}{2}\|\tilde{\beta} - \bar{\beta}\|_2^2 \leq f(\bar{\beta}),$$

   and by local optimality of $\bar{\beta}$, $f(\bar{\beta}) \leq f\left(\tilde{\beta}\right)$. As a result, $\bar{\beta} = \tilde{\beta}$.

2. For large q, $\beta^q \in \operatorname{int}\left(B_\epsilon(\bar{\beta})\right)$, then by Corollary 1, $\nabla F^q(\beta^q) = 0$.

3. Compute [1]

   $$0 = \nabla F^q(\beta^q) = \nabla f(\beta^q) + \sum_{j=1}^{m_1}\left(q g_j^+(\beta^q)\right)\nabla g_j(\beta^q) + \sum_{k=1}^{m_2}\left(q h_k(\beta^q)\right)\nabla h_k(\beta^q) + \beta^q - \bar{\beta}. \tag{6}$$

   For $q = 1, 2, \cdots$, let

   $$\delta^q = \sqrt{1 + \sum_{j=1}^{m_1}\left(q g_j^+(\beta^q)\right)^2 + \sum_{k=1}^{m_2}\left(q h_k(\beta^q)\right)^2},$$

   and $\rho^q = \frac{1}{\delta^q}$, $\lambda_j^q = \frac{q g_j^+(\beta^q)}{\delta^q}$, and $\mu_k^q = \frac{q h_k(\beta^q)}{\delta^q}$ for $j = 1, 2, \cdots, m_1$ and $k = 1, 2, \cdots, m_2$.

[1] Use the fact $\frac{d}{dx}\left(\max\{0, x\}\right)^2 = 2\max\{0, x\}$.

By construction the sequence $\left(\rho^q, \lambda^{q\top}, \mu^{q\top}\right)^\top$ is bounded. By Bolzano-Weierstrass theorem, there exists a subsequence converges to $\left(\rho, \lambda^\top, \mu^\top\right)^\top$ as $q \to \infty$.

4. Let $\mathcal{J} = \{j : \lambda_j > 0\}$, then for large q, $\lambda_j^q \lambda_j > 0$ for $j \in \mathcal{J}$, which, by definition $\lambda_j^q = \frac{q g_j^+(\beta^q)}{\delta^q}$, implies that $g_j(\beta^q) > 0$. By feasibility, $g_j\left(\bar{\beta}\right) \leq 0$, $\beta^q \to \bar{\beta}$ and by continuity of g, we have $g_j\left(\bar{\beta}\right) = 0$ for $j \in \mathcal{J}$.

□

Theorem 3 (Karush–Kuhn–Tucker Necessary Conditions). *Let $\bar{\beta} \in \mathcal{S}$ be a local minimum of (2). Let $\mathcal{J} = \left\{j : g_j\left(\bar{\beta}\right) = 0\right\}$, and suppose that*

$$\left\{\nabla g_j\left(\bar{\beta}\right)\right\}_{j\in\mathcal{J}} \cup \left\{\nabla h_k\left(\bar{\beta}\right)\right\}_{k=1}^{m_2}$$

are linearly independent. Then $\exists\, \lambda \in \mathbb{R}^{m_1}$, $\mu \in \mathbb{R}^{m_2}$ such that

$$\begin{aligned}
&\nabla f\left(\bar{\beta}\right) + \sum_{j=1}^{m_1} \lambda_j \nabla g_j\left(\bar{\beta}\right) + \sum_{k=1}^{m_2} \mu_k \nabla h_k\left(\bar{\beta}\right) = 0, \\
&\lambda_j, \geq 0, j = 1, 2, \cdots, m_1, \\
&\lambda_j g_j\left(\bar{\beta}\right) = 0,\ j = 1, 2, \cdots, m_1.
\end{aligned} \tag{7}$$

Remark 1. The condition

$$\left\{\nabla g_j\left(\bar{\beta}\right)\right\}_{j\in\mathcal{J}} \cup \left\{\nabla h_k\left(\bar{\beta}\right)\right\}_{k=1}^{m_2}$$

are linearly independent is a constraint qualification condition.

A counterexample:

$$\begin{aligned}
\min\quad & x_1 \\
\text{s.t.}\quad & (x_1 - 1)^2 + (x_2 - 1)^2 \leq 1, \\
& (x_1 - 1)^2 + (x_2 + 1)^2 \leq 1.
\end{aligned}$$

The feasible set is a singleton, $\{(1, 0)\}$. The KKT conditions

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2v_1 \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} + 2v_2 \begin{bmatrix} x_1 - 1 \\ x_2 + 1 \end{bmatrix} = \mathbf{0}$$

has no solution at $(1, 0)$.

Theorem 4. *Let $\bar{\beta} \in \mathcal{S}$ be a local minimum of (2), where g_j is convex, $j = 1, 2, \cdots, m_1$ and h_k is affine, $k = 1, 2, \cdots, m_2$. Let $\mathcal{J} = \left\{j : g_j\left(\bar{\beta}\right) = 0\right\}$. Suppose the* **Slater condition** *is satisfied, i.e. $\exists\, \beta' \in \mathcal{S}$ such that $g_j\left(\beta'\right) < 0$ for $j \in \mathcal{J}$. Then $\bar{\beta}$ satisfies the KKT conditions (7).*

- Slater condition with convex constraints form the constraints qualification condition.

Theorem 5. *Let $\bar{\beta} \in \mathcal{S}$ be a local minimum of (2), where g_j is convex, $j = 1, 2, \cdots, m_1$ and h_k is affine, $k = 1, 2, \cdots, m_2$. Then $\bar{\beta}$ satisfies the KKT conditions (7).*

- This implies that KKT conditions are necessary for linearly constrained problems.

- KKT conditions are sufficient for optimality for convex problems.

Theorem 6 (**Karush–Kuhn–Tucker sufficient conditions for convex problems**)**.** *Consider the constrained optimization problem (2), where f, g_j are convex on $\mathcal{B}$, $j = 1, 2, \cdots, m_1$ and h_k is affine, $k = 1, 2, \cdots, m_2$, i.e. it reduces to the convex optimization problem (1). Suppose that $\left(\bar{\beta}, \bar{\lambda}, \bar{\mu}\right)$ satisfies the KKT conditions*

$$\begin{aligned}
&g_j\left(\bar{\beta}\right) \leq 0, j = 1, 2, \cdots, m_1,\\
&h_k\left(\bar{\beta}\right) = 0, k = 1, 2, \cdots, m_2,\\
&\nabla f\left(\bar{\beta}\right) + \sum_{j=1}^{m_1} \bar{\lambda}_j \nabla g_j\left(\bar{\beta}\right) + \sum_{k=1}^{m_2} \bar{\mu}_k \nabla h_k\left(\bar{\beta}\right) = 0,\\
&\bar{\lambda}_j, \geq 0, j = 1, 2, \cdots, m_1,\\
&\bar{\lambda}_j g_j\left(\bar{\beta}\right) = 0,\ j = 1, 2, \cdots, m_1,
\end{aligned} \tag{8}$$

then $\bar{\beta}$ is a global minimum of (2).

Remark 2. Under mild conditions on the functions, the optimality conditions can be generalized to non-differentiable function by replacing gradients with subdifferentials.

1.3 Lagrange Duality

- Consider
$$\begin{aligned}
v_P^* = \min_{\beta \in \mathcal{B}} f\left(\beta\right)\\
\text{s.t. } g_j\left(\beta\right) \leq 0,\ j = 1, 2, \cdots, m_1,\\
h_k\left(\beta\right) = 0,\ k = 1, 2, \cdots, m_2.
\end{aligned} \tag{P}$$
where $\mathcal{B}$ is an open nonempty subset of $\mathbb{R}^p$.

- Define the Lagrangian function $L : \mathbb{R}^p \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$

$$\begin{aligned}
L\left(\beta, \lambda, \mu\right) &= f\left(\beta\right) + \sum_{j=1}^{m_1} \lambda_j g_j\left(\beta\right) + \sum_{k=1}^{m_2} \mu_k h_k\left(\beta\right)\\
&= f(\beta) + \lambda^\top G\left(\beta\right) + \mu^\top H\left(\beta\right).
\end{aligned} \tag{9}$$

- Observe that (P) is equivalent to

$$\inf_{\beta\in\mathcal{B}} \sup_{\lambda\in\mathbb{R}_+^{m_1},\mu\in\mathbb{R}^{m_2}} L\left(\beta,\lambda,\mu\right) \tag{10}$$

  which follows from

$$\sup_{\lambda\in\mathbb{R}_+^{m_1},\mu\in\mathbb{R}^{m_2}} L\left(\beta,\lambda,\mu\right) = \begin{cases} f(\beta) & \text{if } G(\beta)\leq 0, H(\beta)=0 \\ +\infty & \text{otherwise} \end{cases}$$

- For any $\bar{\beta}\in\mathbb{R}^p$, $\left(\bar{\lambda},\bar{\mu}\right)\in\mathbb{R}_+^{m_1}\times\mathbb{R}^{m_2}$,

$$\inf_{\beta\in\mathcal{B}} L\left(\beta,\bar{\lambda},\bar{\mu}\right) \leq L\left(\bar{\beta},\bar{\lambda},\bar{\mu}\right) \leq \sup_{\lambda\in\mathbb{R}_+^{m_1},\mu\in\mathbb{R}^{m_2}} L\left(\bar{\beta},\lambda,\mu\right).$$

- Then

$$\sup_{\lambda\in\mathbb{R}_+^{m_1},\mu\in\mathbb{R}^{m_2}} \inf_{\beta\in\mathcal{B}} L\left(\beta,\lambda,\mu\right) \leq \inf_{\beta\in\mathcal{B}} \sup_{\lambda\in\mathbb{R}_+^{m_1},\mu\in\mathbb{R}^{m_2}} L\left(\beta,\lambda,\mu\right) = v_P^*. \tag{11}$$

- The R.H.S. is equivalent to (P). Define the L.H.S. as the **Lagrangian dual** of (P),

$$v_D^* = \sup_{\lambda\in\mathbb{R}_+^{m_1},\mu\in\mathbb{R}^{m_2}} \psi\left(\lambda,\mu\right), \tag{D}$$

  where

$$\psi\left(\lambda,\mu\right) = \inf_{\beta\in\mathcal{B}} L\left(\beta,\lambda,\mu\right)$$

  is the dual value function.

- The value function $\psi\left(\lambda,\mu\right)$ is the pointwise infimum of affine functions, so it is concave regardless of the convexity of (P).

Theorem 7 (Weak duality). *Let $\bar{\beta}\in\mathbb{R}^p$ be feasible for* (P) *and $\left(\bar{\lambda},\bar{\mu}\right)\in\mathbb{R}^{m_1}\times\mathbb{R}^{m_2}$ be feasible for* (D). *Then we have*

$$\psi\left(\bar{\lambda},\bar{\mu}\right)\leq f\left(\bar{\beta}\right).$$

- **Strong duality**, $v_D^*=v_P^*$, does not hold in general.

Definition 2. $\left(\bar{\beta},\bar{\lambda},\bar{\mu}\right)\in\mathbb{R}^p\times\mathbb{R}^{m_1}\times\mathbb{R}^{m_2}$ is a **saddle point** of the Lagrangian function L associated with (P) if the following conditions hold:

(i) $\bar{\beta}\in\mathcal{B}$.

(ii) $\bar{\lambda} \geq 0$.

(iii) $\forall\, \beta \in \mathcal{B}$, and $(\lambda, \mu) \in \mathbb{R}_+^{m_1} \times \mathbb{R}^{m_2}$, we have

$$L\left(\bar{\beta}, \lambda, \mu\right) \leq L\left(\bar{\beta}, \bar{\lambda}, \bar{\mu}\right) \leq L\left(\beta, \bar{\lambda}, \bar{\mu}\right). \tag{12}$$

Theorem 8. *$\left(\bar{\beta}, \bar{\lambda}, \bar{\mu}\right) \in \mathbb{R}^p \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ is a saddle point of the Lagrangian function L associated with* (P) *if and only if the duality gap between* (P) *and* (D) *is 0 and $\bar{\beta}$ and $\left(\bar{\lambda}, \bar{\mu}\right)$ are the optimal solution to* (P) *and* (D)*, respectively.*

- See (Hastie et al., 2015, Exercise 5.4). Essentially the first inequality in (12) implies feasibility of $\bar{\beta}$ and complementary slackness, which can be used to derive a contradiction, together with the second inequality, if we suppose $\exists\, \tilde{\beta}$ such that $f\left(\tilde{\beta}\right) < f\left(\bar{\beta}\right)$.
- A saddle point implies

  $$\sup_{\lambda \in \mathbb{R}_+^{m_1}, \mu \in \mathbb{R}^{m_2}} \inf_{\beta \in \mathcal{B}} L\left(\beta, \lambda, \mu\right) \geq \psi\left(\bar{\lambda}, \bar{\mu}\right) = f\left(\bar{\beta}\right) \geq \inf_{\beta \in \mathcal{B}} \sup_{\lambda \in \mathbb{R}_+^{m_1}, \mu \in \mathbb{R}^{m_2}} L\left(\beta, \lambda, \mu\right) = v_P^*.$$

  Together with (11), we have

  $$\sup_{\lambda \in \mathbb{R}_+^{m_1}, \mu \in \mathbb{R}^{m_2}} \inf_{\beta \in \mathcal{B}} L\left(\beta, \lambda, \mu\right) = \inf_{\beta \in \mathcal{B}} \sup_{\lambda \in \mathbb{R}_+^{m_1}, \mu \in \mathbb{R}^{m_2}} L\left(\beta, \lambda, \mu\right).$$

Theorem 9 (A special case of Sion's minimax theoreom)**.** *Let L be the Lagrangian function associated with* (P)*. Suppose that*

(i) $\mathcal{B}$ is compact and convex.

(ii) For each $\beta \in \mathcal{B}$, $(\lambda, \mu) \mapsto L\left(\beta, \lambda, \mu\right)$ is continuous and concave on $\mathbb{R}_+^{m_1} \times \mathbb{R}^{m_2}$.

(iii) For each $(\lambda, \mu) \in \mathbb{R}_+^{m_1} \times \mathbb{R}^{m_2}$, $\beta \mapsto L\left(\beta, \lambda, \mu\right)$ is continuous and convex on $\mathcal{B}$.

Then we have

$$\sup_{\lambda \in \mathbb{R}_+^{m_1}, \mu \in \mathbb{R}^{m_2}} \min_{\beta \in \mathcal{B}} L\left(\beta, \lambda, \mu\right) = \min_{\beta \in \mathcal{B}} \sup_{\lambda \in \mathbb{R}_+^{m_1}, \mu \in \mathbb{R}^{m_2}} L\left(\beta, \lambda, \mu\right).$$

- Characterization of saddle points.

Theorem 10 (Saddle point optimality conditions)**.** *$\left(\bar{\beta}, \bar{\lambda}, \bar{\mu}\right) \in \mathbb{R}^p \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ is a saddle point of the Lagrangian function L associated with* (P) *if and only if*

(i) (Primal Feasibility) $\bar{\beta} \in \mathcal{B}$, $G\left(\bar{\beta}\right) \leq 0$, $H\left(\bar{\beta}\right) = 0$.

(ii) *(Dual feasibility)* $\bar{\lambda} \geq 0$.

(iii) *(Lagrangian optimality)* $\bar{\beta} = \arg\min_{\beta \in \mathcal{B}} L\left(\beta, \bar{\lambda}, \bar{\mu}\right)$

(iv) *(Complementary slackness)* $\bar{\lambda}^{\top} G\left(\bar{\beta}\right) = 0$.

- Suppose (P) is reduced to a convex problem (1) where f, g_j are convex and h_j are affine. Then the saddle point optimality conditions are equivalent to the KKT conditions. The strong duality can be retained by the constraint qualification conditions, Slater's condition for example.

2 Numerical Algorithms

2.1 Gradient Descent

2.1.1 Unconstrained Gradient Descent

- Gradient descent is an iterative algorithm for solving $\nabla f\left(\beta^*\right) = 0$

- Iterate

$$\beta^{t+1} = \beta^t - s^t \nabla f\left(\beta^t\right), \tag{13}$$

  for $t = 0, 1, 2, \ldots$, where $s^t > 0$ is a step size parameter.

- Direction of descent: $-\nabla f\left(\beta^t\right)$

- In general, the class of **descent methods** is based on choosing a direction $\Delta^t \in \mathbb{R}^p$ such that $\langle \nabla f\left(\beta^t\right), \Delta^t \rangle < 0$, then update

$$\beta^{t+1} = \beta^t + s^t \Delta^t \quad \text{for } t = 0, 1, 2, \ldots$$

- **Newton's method** for twice continuously differentiable functions,

$$\Delta^t = -\left(\nabla^2 f\left(\beta^t\right)\right)^{-1} \nabla f\left(\beta^t\right)$$

  - a Newton step (with stepsize one) amounts to exactly minimizing the second-order Taylor approximation
  - quadratic rate of convergence
  - Computing the Hessian matrix is expensive for large-scale problems.

- **Quasi-Newton methods** approximate the Hessian matrix by a positive definite matrix B^t, then update

$$\Delta^t = -\left(B^t\right)^{-1} \nabla f\left(\beta^t\right)$$

- Choose step-size
  - Exact line search: $s^t = \arg\min_s f\left(\beta^t + s\Delta^t\right)$
  - Limited minimization rule: $s^t = \underset{s\in[0,1]}{\arg\min} f\left(\beta^t + s\Delta^t\right)$
  - Backtracking line search: Given parameters $\alpha \in (0, 0.5]$ and $\gamma \in (0,1)$ and an initial stepsize $s = 1$, repeat $s \leftarrow \gamma s$ until the Armijo condition

$$f\left(\beta^t + s\Delta^t\right) \leq f\left(\beta^t\right) + \alpha s\left\langle \nabla f\left(\beta^t\right), \Delta^t\right\rangle$$

    is satisfied.

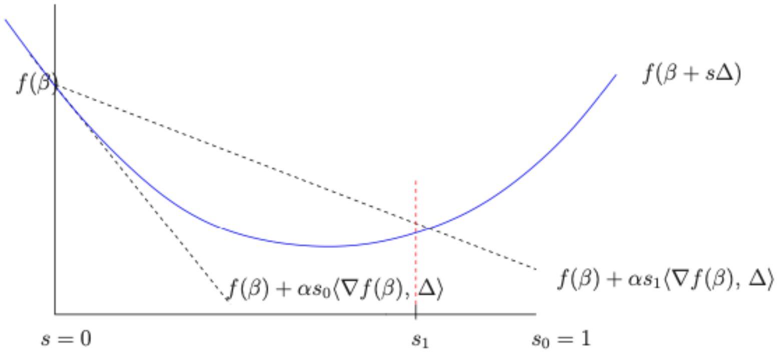


Figure 5.4 *Armijo rule or backtracking line search. Starting with step-size $s_0 = 1$, we repeatedly reduce s by a fraction γ until the condition $f(\beta + s\Delta) \leq f(\beta) + \alpha s\langle\nabla f(\beta), \Delta\rangle$ is satisfied. This is achieved here at s_1.*

2.1.2 Projected Gradient Descent

- Problems that involve additional side constraints

- Gradient descent (13) can be viewed as the combination of a local linear approximation to f combined with a quadratic smoothness term

$$\beta^{t+1} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ f\left(\beta^t\right) + \left\langle \nabla f\left(\beta^t\right), \beta - \beta^t\right\rangle + \frac{1}{2s^t}\left\|\beta - \beta^t\right\|_2^2 \right\}. \tag{14}$$

- Projected gradient descent

$$\beta^{t+1} = \arg\min_{\beta \in \mathcal{C}} \left\{ f\left(\beta^t\right) + \left\langle \nabla f\left(\beta^t\right), \beta - \beta^t \right\rangle + \frac{1}{2s^t} \left\| \beta - \beta^t \right\|_2^2 \right\}.$$

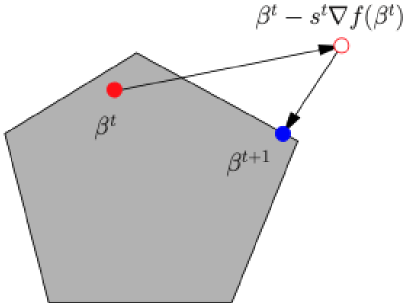


Figure 5.5 *Geometry of projected gradient descent. Starting from the current iterate β^t, it moves in the negative gradient direction to $\beta^t - s^t \nabla f(\beta^t)$, and then performs a Euclidean projection of the result back onto the convex constraint set $\mathcal{C}$ in order to obtain the next iterate β^{t+1}.*

2.1.3 Proximal Gradient Method

- The objective function can be decomposed as a convex and differentiable component g and a convex but nondifferentiable component h,

$$\min_{\beta} f\left(\beta\right) = g\left(\beta\right) + h\left(\beta\right)$$

- Cannot use gradient methods; Do not want to use subgradient methods due to speed concerns

- If h has the property that the **proximal mapping (prox-operator)** of h is easy to compute, we can use the proximal gradient method

- The proximal mapping (prox-operator) of a convex function h is defined as

$$\operatorname{prox}_h(x) := \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|x - \theta\|_2^2 + h(\theta) \right\} \tag{15}$$

  - $h(x) = 0$: $\operatorname{prox}_h(x) = x$
  - $h(x) = I_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise} \end{cases}$, $\operatorname{prox}_h(x) = \operatorname*{argmin}_{u \in C} \|u - x\|_2^2 = P_C(x)$

- $h(x) = \|x\|_1$, then $\text{prox}_h(x)_i = \begin{cases} x_i - 1 & x_i \geq 1 \\ 0 & |x_i| \leq 1 \\ x_i + 1 & x_i \leq 1 \end{cases}$
- $f(x) = \|x\|_2$, then $\text{prox}_{tf}(x) = \begin{cases} \left(1 - \frac{t}{\|x\|_2}\right) x & \|x\|_2 \geq t \\ 0 & \text{otherwise} \end{cases}$
- $f(x) = \frac{1}{2}x^\top Ax + b^\top x + c, \quad \text{prox}_{tf}(x) = (I + tA)^{-1}(x - tb)$ with $A \succ 0$.
- If h is closed and convex, then $\text{prox}_h(x)$ exists (closed and bounded sublevel sets) and is unique (strong convexity)
- **Subgradient characterization**:

  $$u = \text{prox}_h(x) \Leftrightarrow x - u \in \partial h(u) \Leftrightarrow h(z) \geq h(u) + (x-u)^\top(z-u), \forall z$$

  which can be derived from the first order condition.
- $u = P_C(x) \Leftrightarrow (x-u)^\top(z-u) \leq 0; \forall z \in C$
- **Nonexpansiveness (Lipschitz continuous with $L = 1$)** : if $u = \text{prox}_h(x)$, $v = \text{prox}_h(y)$, then

  $$(u-v)^\top(x-y) \geq \|u-v\|_2^2 \Rightarrow \|x-y\|_2 \geq \|u-v\|_2$$

- **Moreau decomposition**:

  $$x = \text{prox}_f(x) + \text{prox}_{f^*}(x) \,\forall x$$

  where $f^*(y) = \sup_{x \in \text{dom}\, f}\left(y^\top x - f(x)\right)$ is the **conjugate** of f, which is convex regardless of the convexity of f.

- Recall (14) and consider the generalized gradient update

$$\beta^{t+1} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ g\left(\beta^t\right) + \left\langle \nabla g\left(\beta^t\right), \beta - \beta^t \right\rangle + \frac{1}{2s^t}\left\|\beta - \beta^t\right\|_2^2 + h(\beta) \right\}$$

- The update rule is equivalent to

$$\beta^{t+1} = \text{prox}_{s^t h}\left(\beta^t - s^t \nabla g\left(\beta^t\right)\right)$$

((Hastie et al., 2015, Exercise 5.7)) This can be easily shown by the subgradient characterization

- Projected gradient descent in the form of proximal mapping,

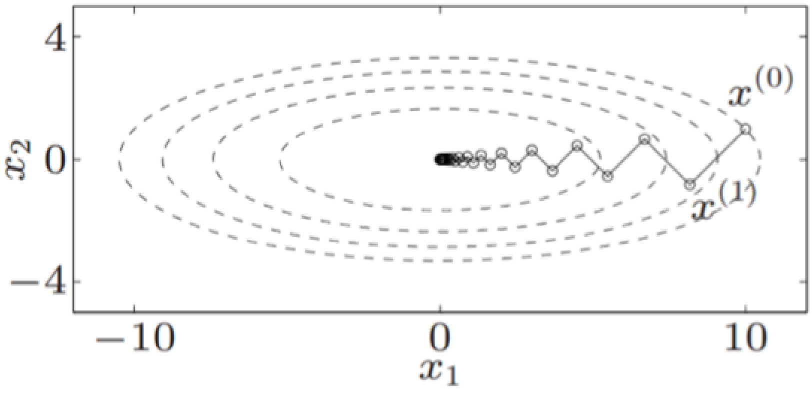
$$\beta^{t+1} = \operatorname{prox}_{I_C}\left(\beta^t - s^t \nabla g\left(\beta^t\right)\right)$$

- Example (Lasso). $h\left(\theta\right) = \lambda\left|\theta\right|_1$ and $g(\beta) = \frac{1}{2N}\|y - X\beta\|_2^2$. Then proximal gradient update

$$\beta^{t+1} = \mathcal{S}_{s^t\lambda}\left(\beta^t + s^t \frac{1}{N} X^\top\left(y - X\beta^t\right)\right)$$

2.1.4 Accelerated Gradient Methods

- Gradient descent exhibit ”zig-zagging” behavior dependent on the objective functions. For example, $f(x) = (1/2)\left(x_1^2 + \gamma x_2^2\right)$ with large γ .



- Nesterov (2007) proposed the class of accelerated gradient methods that use weighted combinations of the current and previous gradient directions.

- Initialize $\beta^0 = \theta^0$. Update according to

$$\begin{aligned}\beta^{t+1} &= \theta^t - s^t \nabla f\left(\theta^t\right)\\ \theta^{t+1} &= \beta^{t+1} + \frac{t}{t+3}\left(\beta^{t+1} - \beta^t\right)\end{aligned}$$

- Replace the gradient update step if there is nondifferentiable h evolved,

$$\beta^{t+1} = \operatorname{prox}_{s^t h}\left(\theta^t - s^t \nabla g\left(\theta^t\right)\right)$$

2.2 Coordinate Descent

- **Coordinate descent** is an iterative algorithm that updates from β^t to β^{t+1} by choosing a single coordinate to update, and then performing a univariate minimization over this coordinate.

- If coordinate k is chosen,

$$\beta_k^{t+1} = \underset{\beta_k}{\arg\min} f\left(\beta_1^t, \beta_2^t, \ldots, \beta_{k-1}^t, \beta_k, \beta_{k+1}^t, \ldots, \beta_p^t\right)$$

- A sufficient condition for convergence: the objective function is continuously differentiable and strongly convex in each coordinate.

- Separability condition: f has the additive decomposition $f\left(\beta_1, \ldots \beta_p\right) = g\left(\beta_1, \ldots \beta_p\right) + \sum_{j=1}^p h_j\left(\beta_j\right)$

  - g is convex and differentiable and h_j is convex and possibly nondifferentiable.
  - Example: Lasso

- Tseng (2001) gives a more general and intuitive condition for convergence of CD

  Define

$$f'(\beta;\Delta) := \liminf_{s\downarrow 0} \frac{f(\beta+s\Delta)-f(\beta)}{s}$$

  regularity condition If $f'\left(\beta; e^j\right) \geq 0$, $\forall, j = 1, 2, \cdots, p$, then $f'(\beta;\Delta) \geq 0$, $\forall\, \Delta \in \mathbb{R}^p$.

  - Separability implies regularity.
  - An example without separability but with regularity:

$$h\left(\beta_1, \ldots, \beta_p\right) = |\beta|^\top P|\beta| = \sum_{j,k=1}^p \left|\beta_j\right| P_{jk}\left|\beta_k\right|$$

    where $P \succeq 0$.

  - An example that fails regularity: Fused Lasso. Here the non-differentiable component takes the form $h(\beta) = \sum_{j=1}^p |\beta_j - \beta_{j-1}|$.

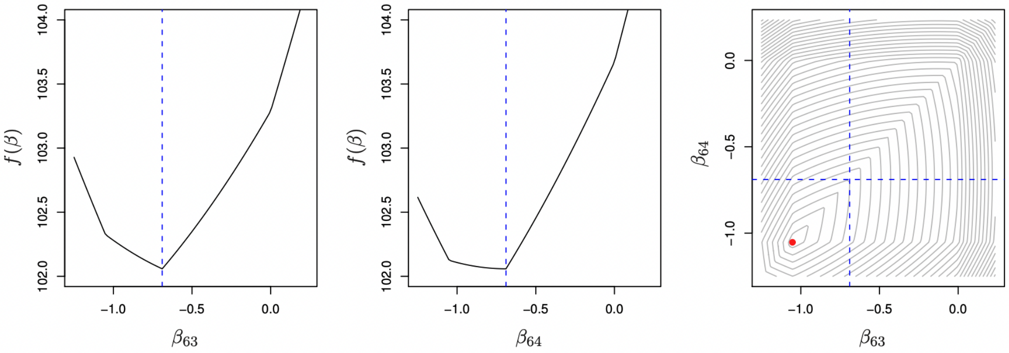


Figure 5.8 *Failure of coordinate-wise descent in a fused lasso problem with 100 parameters. The optimal values for two of the parameters, β_{63} and β_{64}, are both -1.05, as shown by the dot in the right panel. The left and middle panels show slices of the objective function f as a function of β_{63} and β_{64}, with the other parameters set to the global minimizers. The coordinate-wise minimizer over both β_{63} and β_{64} (separately) is -0.69, rather than -1.05. The right panel shows contours of the two-dimensional surface. The coordinate-descent algorithm is stuck at the point $(-0.69, -0.69)$. Despite being strictly convex, the surface has corners, in which the coordinate-wise procedure can get stuck. In order to travel to the minimum we have to move both β_{63} and β_{64} together.*

- **Example (Lasso)**. The optimality condition,

$$-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{k=1}^{p}x_{ik}\beta_k\right)x_{ij}+\lambda s_j=0$$

where $s_j \in \operatorname{sign}(\beta_j)$.

Define partial residual $r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\widehat{\beta}_k$, the solution,

$$\widehat{\beta}_j = \frac{\mathcal{S}_\lambda\left(\frac{1}{N}\sum_{i=1}^{N} r_i^{(j)} x_{ij}\right)}{\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2}.$$

where $\mathcal{S}_\lambda(\theta) = \operatorname{sign}(\theta)(|\theta|-\lambda)_+$ is the soft-thresholding operator.

For centered and standardized data,

$$\widehat{\beta}_j = \mathcal{S}_\lambda\left(\tilde{\beta}_j\right)$$

where $\tilde{\beta}_j$ is the solution of the univariate regression problem of partial residual $r_i^{(j)}$ on x_{ij}.

Efficiency Improvement:

- (*Naive updating*) Note that

  $$\frac{1}{N}\sum_{i=1}^{N} x_{ij} r_i^{(j)} = \frac{1}{N}\sum_{i=1}^{N} x_{ij} r_i + \widehat{\beta}_j.$$

  Computing the inner product $\langle \mathbf{x}_j, \mathbf{r} \rangle$ is $\mathcal{O}(N)$, and updating r_i is also $\mathcal{O}(N)$. A full cycle through all p variables costs $\mathcal{O}(pN)$.

- (*Covariance updating*) $\langle x_j, r \rangle$ can be further decomposed as

  $$\sum_{i=1}^{N} x_{ij} r_i = \langle \mathbf{x}_j, \mathbf{y} \rangle - \sum_{k \| \widehat{\beta}_k | > 0} \langle \mathbf{x}_j, \mathbf{x}_k \rangle \widehat{\beta}_k.$$

  $\langle \mathbf{x}_j, \mathbf{y} \rangle$ can be computed offline. $\langle \mathbf{x}_j, \mathbf{x}_k \rangle$ can be computed whenever $\mathbf{x}_k$ enters the model. If no new variable enters, $\mathcal{O}(p)$; otherwise, $\mathcal{O}(pN)$. This is efficiency when $N \gg p$.

- (*Warm Starts*) For (decreasing) sequence $\{\lambda_0\}_{\ell=0}^{L}$ with

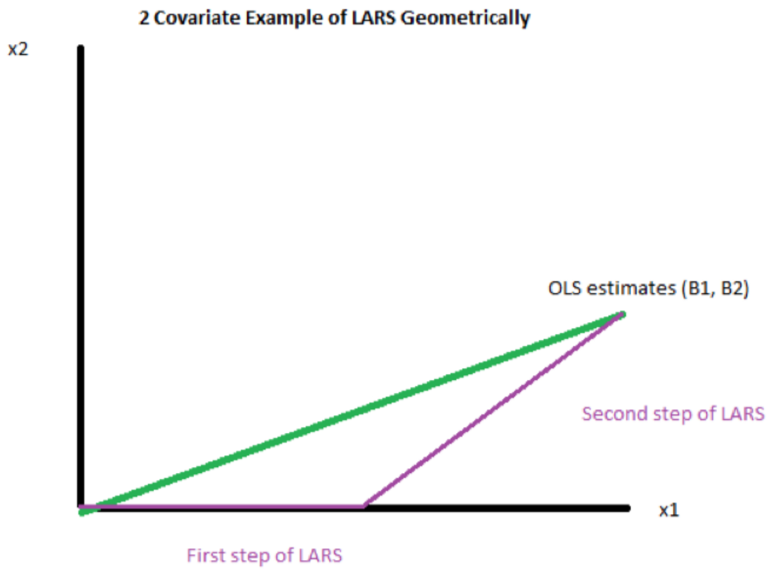
  $$\lambda_0 = \frac{1}{N}\max_j |\langle \mathbf{x}_j, \mathbf{y} \rangle|,$$

  solution $\widehat{\beta}(\lambda_\ell)$ is typically a good starting point for $\widehat{\beta}(\lambda_{\ell+1})$.

  `glmnet` generate a geometric sequence from λ_0 to $\lambda_L = \epsilon\lambda_0$ with ϵ and L specified by the user.

  ```
  lambda_min <- lambda_min_ratio * lambda_max
  lambda_seq <- exp(seq(log(lambda_max), log(lambda_min), length.out = nlambda))
  ```

  - Cross-validation based on $\widehat{\lambda}(\lambda_\ell)$
  - A modified least angle regression (LAR) algorithm can generate the solution path of Lasso. Details see (Hastie et al., 2015, Section 5.6).



- (*Active-set convergence*) Find the active set $\mathcal{A}$ after one iteration start from $\widehat{\beta}\left(\lambda_{\ell-1}\right)$. Upon convergence with only $\mathcal{A}$, check $\frac{1}{N}\left|\langle \mathbf{x}_j, \boldsymbol{r}\rangle\right| < \lambda_\ell$ for omitted variables.
- (*Strong set convergence*) Define

$$\mathcal{S} = \left\{ j : \left|\frac{1}{N}\left\langle \mathbf{x}_j, \boldsymbol{r}\right\rangle\right| > \lambda_\ell - \left(\lambda_{\ell-1} - \lambda_\ell\right) \right\}$$

  and then restrict attention only variables in $\mathcal{S}$.

The algorithm can be generalized to have heterogeneous penalty weights,

$$\lambda \sum_{j=1}^{p} \gamma_j P_\alpha\left(\beta_j\right).$$

Adaptive Lasso can be implemented by `glmnet` with a weight argument specified.

```
glmnet(x, y, lambda = lambda * sum(w) / p, penalty.factor = w)
```

- **Example (Logistic Regression)** The log-likelihood,

$$\ell\left(\beta_0, \beta\right) = \frac{1}{N} \sum_{i=1}^{N} \left[y_i \cdot \left(\beta_0 + x_i^T \beta\right) - \log\left(1 + e^{\beta_0 + x_i^T \beta}\right) \right].$$

Given the current estimate $\left(\widetilde{\beta}_0, \widetilde{\beta}\right)$, we derive the second order Taylor approximation around

$\left(\widetilde{\beta}_0, \widetilde{\beta}\right)$,

$$\ell_Q\left(\beta_0, \beta\right) = -\frac{1}{2N}\sum_{i=1}^{N} w_i\left(z_i - \beta_0 - x_i^T\beta\right)^2 + C\left(\widetilde{\beta}_0, \widetilde{\beta}\right),$$

where C is a constant, $\tilde{p}\left(x_i\right) = p\left(x_i; \widetilde{\beta}_0, \widetilde{\beta}\right)$, $w_i = \tilde{p}\left(x_i\right)\left(1 - \tilde{p}\left(x_i\right)\right)$, and $z_i = \widetilde{\beta}_0 + x_i^T\widetilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1-\tilde{p}(x_i))}$.

The Newton update is obtained by minimizing ℓ_Q, which is a simple weighted least square.

Apply the coordinate descent to the quadratic approximaiton with regularization,

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left\{-\ell_Q\left(\beta_0, \beta\right) + \lambda P_\alpha(\beta)\right\}.$$

- Speed comparison:

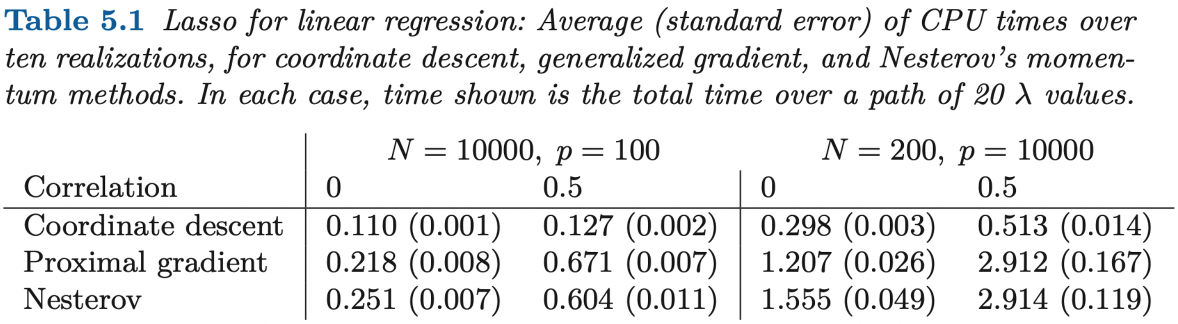


Table 5.1 *Lasso for linear regression: Average (standard error) of CPU times over ten realizations, for coordinate descent, generalized gradient, and Nesterov's momentum methods. In each case, time shown is the total time over a path of 20 λ values.*

	$N = 10000,\ p = 100$		$N = 200,\ p = 10000$	
Correlation	0	0.5	0	0.5
Coordinate descent	0.110 (0.001)	0.127 (0.002)	0.298 (0.003)	0.513 (0.014)
Proximal gradient	0.218 (0.008)	0.671 (0.007)	1.207 (0.026)	2.912 (0.167)
Nesterov	0.251 (0.007)	0.604 (0.011)	1.555 (0.049)	2.914 (0.119)

2.3 Augmented Lagrangian Method

- Consider

$$\min_{\beta\in\mathbb{R}^p} f(\beta) \quad \text{subject to } \mathbf{A}\beta = c$$

where $\mathbf{A} \in \mathbb{R}^{d\times p}$.

- Gradient projection method

$$\beta^{t+1} = \arg\min_{\beta:\mathbf{A}\beta=c} \left\|\beta - \left(\beta^t - s^t\nabla f\left(\beta^t\right)\right)\right\|_2^2.$$

- Projection onto the affine set $\mathbf{A}\beta = c$,

$$P_{\{\beta:\mathbf{A}\beta=c\}}(u) = \arg\min_{\beta:\mathbf{A}\beta=c} \|\beta - u\|_2^2 = u + \mathbf{A}^\top\left(\mathbf{A}\mathbf{A}^\top\right)^{-1}(c - \mathbf{A}u),$$

can be expensive unless $d \ll p$, or $\mathbf{A}\mathbf{A}^\top = I$.

- Define the augmented Lagrangian function

$$L_\rho\left(\beta,\mu\right)=f\left(\beta\right)+\mu^\top\left(\mathbf{A}\beta-c\right)+\frac{\rho}{2}\left\|\mathbf{A}\beta-c\right\|_2^2.$$

  The augmented Lagrangian method (ALM) is (starting with $\mu^0=0$),

$$\begin{aligned}\beta^{t+1}&=\arg\min_\beta L_\rho\left(\beta,\mu^t\right)\\ \mu^{t+1}&=\mu^t+\rho\left(\mathbf{A}\beta^{t+1}-c\right).\end{aligned}$$

- ALM method is the proximal point method applied to the dual problem.
  - Proximal point method

$$\beta^{t+1}=\operatorname{prox}_{s^tf}\left(\beta^t\right)=\arg\min_u\left(f(u)+\frac{1}{2s^t}\left\|u-\beta^t\right\|_2^2\right)$$

  - The Lagrangian function is

$$L\left(\beta,\mu\right)=f\left(\beta\right)+\mu^\top\left(\mathbf{A}\beta-c\right)$$

  - Dual problem is

$$\max_\mu\inf_\beta L\left(\beta,\mu\right)=\max_\mu\left\{-f^*\left(-\mathbf{A}^\top\mu\right)-c^\top\mu\right\}=\min_\mu\left\{f^*\left(-\mathbf{A}^\top\mu\right)+c^\top\mu\right\}$$

  - Let $h\left(\mu\right)=f^*\left(-\mathbf{A}^\top\mu\right)+c^\top\mu$,

$$\operatorname{prox}_{\rho h}\left(\mu\right)=\arg\min_a\left(h(a)+\frac{1}{2\rho}\left\|a-\mu\right\|_2^2\right)$$

    Optimality condition:

$$\frac{1}{\rho}\left(a-\mu\right)+c\in\mathbf{A}\partial f^*\left(-\mathbf{A}^\top a\right)$$

  - Write the augmented Lagrangian as

$$\begin{aligned}&\arg\min_\beta f\left(\beta\right)+\mu^\top\left(\mathbf{A}\beta-c\right)+\frac{\rho}{2}\left\|\mathbf{A}\beta-c\right\|_2^2\\ =&\arg\min_\beta\left\{f\left(\beta\right)+\frac{\rho}{2}\left\|\mathbf{A}\beta-c+\frac{\mu}{\rho}\right\|_2^2\right\}\end{aligned}$$

Optimality condition:

$$\begin{aligned} & -\mathbf{A}^\top\left[\mu+\rho\left(A\beta-c\right)\right] \in \partial f\left(\beta\right) \\ \rightarrow \quad & \beta \in \partial f^*\left(-\mathbf{A}^\top\left[\mu+\rho\left(A\beta-c\right)\right]\right) \\ \rightarrow_{a=\mu+\rho(A\beta-c)} \quad & \frac{1}{\rho}\left(a-\mu\right) \in A\beta \in A\partial f^*\left(-\mathbf{A}^\top a\right) \end{aligned}$$

- Hence

$$\operatorname{prox}_{\rho h}\left(\mu\right)=\mu+\rho\left(A\hat{\beta}-c\right)$$

  where $\hat{\beta}=\arg\min_\beta L_\rho\left(\beta,\mu\right)$.

- Three variants of L_1-regularized problems,
  - Lasso

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{2N}\|\mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1$$

    Default algorithm: Proximal gradient method
  - Constrained form

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{2N}\|\mathbf{y}-\mathbf{X}\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_1\le t$$

    Default algorithm: Projected gradient method
  - Basis pursuit

$$\min_\beta\|\beta\|_1 \quad \text{s.t.} \quad \mathbf{X}\beta=\mathbf{y}$$

    Default algorithm: Augmented Lagrangian method

2.4 Alternating Direction Method of Multipliers (ADMM)

- Comprehensive survey: Boyd et al. (2011)
- Consider the problem,

$$\min_{\beta\in\mathbb{R}^m,\theta\in\mathbb{R}^n} f(\beta)+g(\theta) \quad \text{subject to } \mathbf{A}\beta+\mathbf{B}\theta=c$$

  where f and g are convex, $\mathbf{A}\in\mathbb{R}^{d\times m}$ and $\mathbf{B}\in\mathbb{R}^{d\times n}$.

- A special case:

$$\min_{\beta\in\mathbb{R}^m,\theta\in\mathbb{R}^n} f(\beta)+g(\beta)$$

  by letting $\theta=\beta$.

- Augmented Lagrangian,

$$L_\rho(\beta,\theta,\mu) := f(\beta) + g(\theta) + \langle \mu, \mathbf{A}\beta + \mathbf{B}\theta - c\rangle + \frac{\rho}{2}\|\mathbf{A}\beta + \mathbf{B}\theta - c\|_2^2$$

- ADMM

$$\begin{aligned}
\beta^{t+1} &= \underset{\beta\in\mathbb{R}^m}{\arg\min} L_\rho\left(\beta,\theta^t,\mu^t\right)\\
\theta^{t+1} &= \underset{\theta\in\mathbb{R}^n}{\arg\min} L_\rho\left(\beta^{t+1},\theta,\mu^t\right)\\
\mu^{t+1} &= \mu^t + \rho\left(\mathbf{A}\beta^{t+1} + \mathbf{B}\theta^{t+1} - c\right),
\end{aligned}$$

- **Example (Lasso)**

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\theta\|_1 \quad \text{s.t.} \quad \beta - \theta = 0$$

  The ADMM updates

$$\begin{aligned}
\beta^{t+1} &= \left(\mathbf{X}^T\mathbf{X} + \rho\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} + \rho\theta^t - \mu^t\right)\\
\theta^{t+1} &= \mathcal{S}_{\lambda/\rho}\left(\beta^{t+1} + \mu^t/\rho\right)\\
\mu^{t+1} &= \mu^t + \rho\left(\beta^{t+1} - \theta^{t+1}\right)
\end{aligned}$$

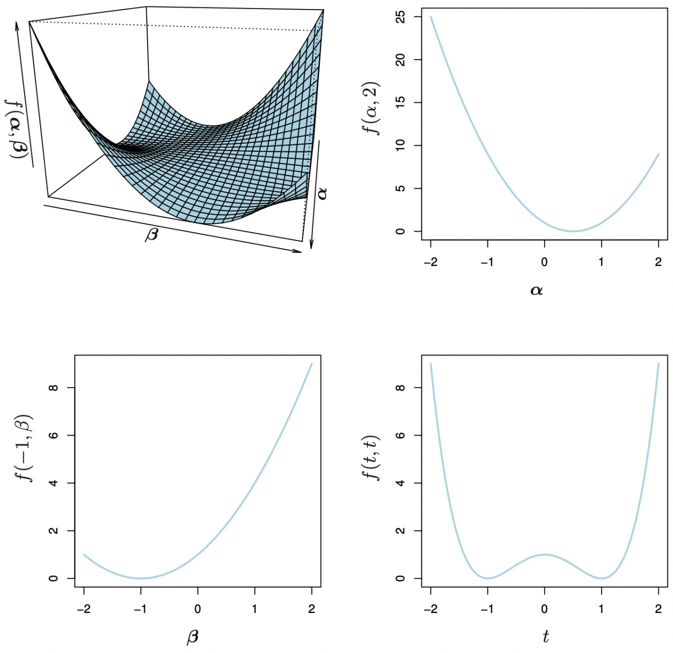
  The Ridge regression update for β requires an initial SVD of $\mathbf{X}$ implemented offline which requires $\mathcal{O}\left(p^3\right)$ operations. Subsequent iterations cost $\mathcal{O}\left(Np\right)$, which is similar to coordinate descent.

2.5 Minorization-Maximization Algorithms

- The objective functin f is possibly nonconvex
- A function Ψ majorizes f at β if $f(\beta) \leq \Psi(\beta,\theta)$ for all $\theta \in \mathbb{R}^p$ with equality holds when $\beta = \theta$.
- MM algorithm: $\beta^{t+1} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\Psi\left(\beta,\beta^t\right)$, $t = 0, 1, 2, \cdots$
- $f\left(\beta^t\right) = \Psi\left(\beta^t,\beta^t\right) \geq \Psi\left(\beta^{t+1},\beta^t\right) \geq f\left(\beta^{t+1}\right)$

2.6 Biconvexity and Alternating Minimization

- A function $f\left(\alpha,\beta\right)$ is biconvex if $\alpha \mapsto f(\alpha,\beta)$ is convex and $\beta \mapsto f(\alpha,\beta)$ is also convex.
- Example: $f(\alpha,\beta) = (1-\alpha\beta)^2$ for $|\alpha| \leq 2, |\beta| \leq 2$



- Alternate Convex Search (ACS)

  (a) Initialize (α^0, β^0) at some point in $\mathcal{C}$

  (b) For iterations $t = 0, 1, 2, \cdots,$

  (i) Fix $\beta = \beta^t$, update $\alpha^{t+1} = \arg\min_{\alpha \in \mathcal{C}_{\beta^t}} f(\alpha, \beta^t)$

  (ii) Fix $\alpha = \alpha^{t+1}$, update $\beta^{t+1} = \arg\min_{\beta \in \mathcal{C}_{\alpha^{t+1}}} f(\alpha^{t+1}, \beta)$

- Only objective values converge to some limit point. The solution sequence may not converge.

  $(\alpha^*, \beta^*) \in \mathcal{C}$ is a partial optimum if

$$
\begin{aligned}
f(\alpha^*, \beta^*) &\leq f(\alpha^*, \beta) \quad \text{for all} \quad \beta \in \mathcal{C}_{\alpha^*} \\
f(\alpha^*, \beta^*) &\leq f(\alpha, \beta^*) \quad \text{for all} \quad \alpha \in \mathcal{C}_{\beta^*}
\end{aligned}
$$

- Example (Mixed membership in Grouped panel data models).

$$
\min_{\theta, \alpha_{gt}, \gamma_{ig}} \; Q(\theta, \alpha, \gamma) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x'_{it}\beta - \sum_{g} \alpha_{tg}\gamma_{ig})^2
$$

$$
\text{s.t. } \sum_{g} \gamma_{ig} = 1, \; \forall i, \gamma_{ig} \in [0, 1], \; \forall i, g
$$

3 Implementation with Optimization Solvers

- R's optimization infrastructure has been constantly improving. R Optimization Task View gives a survey of the available CRAN packages.

- De facto R package: optimx

- General purpose nonlinear programming solver NLOPT

  Example (OLS).

```
f <- function(b, y, x){sum((y - x %*% b)^2)}
s <- function(b, y, x){-2 * t(x) %*% (y - x %*% b)}
opts <- list(algorithm = "NLOPT_LD_SLSQP", xtol_rel = 0.0001)
res_nloptr <- nloptr::nloptr(x0 = c(0, 0, 0), eval_f = f,
                             eval_grad_f =s, opts = opts, y = y, x = x)
```

  NLOPT algorithms:

  - Nelder-Mead (NLOPT_LN_NELDERMEAD): A derivative-free simplex method. It searches a local minimum by reflection, expansion and contraction.
  - BFGS (NLOPT_LD_LBFGS): A quasi-Newton algorithm.
  - SQP (NLOPT_LD_SLSQP) : A sequential quadratic programming (SQP) algorithm for nonlinearly constrained gradient-based optimization.
  - For more options, see NLOPT Algorithms.

- Disciplined convex programming solver CVX

- Commercial optimization solver MOSEK specialized for convex problems

- Commercial optimization solver Gurobi specialized for convex problems and integer programming

3.1 Example: Lasso

- Recall the standard Lasso problem is of the form

$$\min_{\beta} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

- Use CVXR, the R wrapper of the CVX solver, to solve the Lasso problem.

```r
# CVXR
beta <- Variable(p)
obj <- sum_squares(y - x %*% beta) / (2 * n) + lambda * p_norm(beta, 1)
prob <- Problem(Minimize(obj))
result <- solve(prob, verbose = FALSE)
```

```
## Time difference of 0.2630119 secs[2]
```

```r
result <- solve(prob, solver = "MOSEK", verbose = FALSE)
```

```
## Time difference of 0.09962988 secs
```

- Since the Lasso problem is well-studied and we know the problem is convex, it is desirable to directly call MOSEK to skip the convexity check steps. To invoke MOSEK, we need to transform the problem to a standard form of conic programming that the solver can recognize.
- We first deal with $\|\beta\|_1$.
  - The $p \times 1$ vector β can be decomposed into a positive part $\beta^+ = (\max\{0, \beta_j\})_{j=1}^p$ and a negative part $\beta^- = (\max\{0, -\beta_j\})_{j=1}^p$, so that $\beta = \beta^+ - \beta^-$ and $\|\beta\|_1 = e'\beta^+ + e'\beta^-$, where e is the $p \times 1$ vector with all elements equal to 1.
- Next, we transform the l_2-norm $\|y - X\beta\|_2^2$ to a second-order conic constraint.
  - Consider a minimization problem with $\|\nu\|_2^2$ in the objective function. We can use a new parameter t to replace it and add a conic constraint $\|\nu\|_2^2 \leq t$, which is equivalent to $\left\|\left(\nu, \frac{t-1}{2}\right)\right\|_2 \leq \frac{t+1}{2}$.
  - Thus we obtain a standard conic constraint $\|(\nu, s)\|_2 \leq r$, where $s = \frac{t-1}{2}$ and $r = \frac{t+1}{2}$.
- We rewrite the Lasso problem as

$$\min_{\theta} \lambda\left(e'\beta^+ + e'\beta^-\right) + \frac{t}{n} \quad \text{s.t.} \quad \nu = y - X\left(\beta^+ - \beta^-\right), \ \|(v, s)\|_2 \leq r, \ s = \frac{t-1}{2}, \ r = \frac{t+1}{2}$$

  where $\theta = (\beta^+, \beta^-, \nu, t, s, r)$. This problem is of the standard form of second-order conic programming.

[2] Use simulated data with $n = 100$ and $p = 20$.

- In matrix notation, the Lasso problem is

$$\min_{\theta} \lambda\left(e'\beta^{+} + e'\beta^{-}\right) + \frac{t}{n}$$

$$\text{s.t.} \quad \begin{bmatrix} X & -X & I_n & \mathbf{0}_{n\times 3} \\ & \mathbf{0}_{2\times(n+2p)} & & \begin{matrix} -\frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{matrix} \end{bmatrix} \theta = \begin{bmatrix} y \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \ \|(v,s)\|_2 \leq r, \ \beta^{+}, \beta^{-} \geq 0$$

where the inequality for a vector is taken elementwisely. The following annotated `R` code snippet implements the matrix form.

```
P = list(sense = "min")

# Linear coefficients in objective
P$c = c(rep(lambda, 2*p), rep(0, n), 1/n, 0, 0)

# The matrix in linear constraints
A = as.matrix.csr(X)
A = cbind(A, -A, as(n, "matrix.diag.csr"), as.matrix.csr(0, n, 3))
A = rbind(A, cbind(as.matrix.csr(0, 2, 2*p + n),
                                 as.matrix.csr(c(-.5, -.5, 1, 0, 0, 1), 2, 3)))
P$A = as(A,"CsparseMatrix")

# Right-hand side of linear constraints
P$bc = rbind(c(y, -0.5, 0.5), c(y, -0.5, 0.5))

# Constraints on variables
P$bx = rbind(c(rep(0, 2 * p), rep(-Inf, n), rep(0, 3)), c(rep(Inf, 2*p+n+3)))

# Conic constraints
P$cones = matrix(list("QUAD", c(n+2*p+3, (2*p+1):(2*p+n), n+2*p+2)), 2, 1)
rownames(P$cones) = c("type", "sub")

result = mosek(P, opts = list(verbose = verb))
xx = result$sol$itr$xx
coef = xx[1:p] - xx[(p+1):(2*p)]
```

```
## Time difference of 0.006147146 secs
```

```r
glmnet(x, y, lambda = lambda)
```

```
## Time difference of 0.003031015 secs
```

3.2 Additional Example: Exponential/Logarithm Formulation

In limited dependent variable models it is common to see exponential, logarithm or power terms in objective functions. These nonlinear functions can be easily formulated with power cones and exponential cones.

Example 1 (Poisson maximum likelihood estimator)**.** The Poisson maximum likelihood estimator is defined as

$$\min_{\beta} -\frac{1}{n}\sum_{i=1}^{n}\left(y_i x_i'\beta - \exp\left(x_i'\beta\right)\right)$$

where $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ are observed data, and β is the parameter of interests. This optimization problem involves the component $\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$, which is non-separable. Define $v_i = x_i'\beta$, and the objective becomes

$$\min_{v,\beta} \frac{1}{n}\sum_{i=1}^{n}\left(-y_i v_i + \exp\left(v_i\right)\right).$$

We introduce the auxiliary variable t_i to replace $\exp\left(v_i\right)$ in the objective and the constraint $t_i \geq \exp\left(v_i\right)$ is equivalent to $\left(t_i, 1, v_i\right) \in \mathcal{K}_{\text{exp}}$. The estimator can be then formulated as

$$\min_{t,v,\beta} \frac{1}{n}\sum_{i=1}^{n}\left(-y_i v_i + t_i\right) \quad \text{s.t. } v_i = x_i'\beta,\ \left(t_i, 1, v_i\right) \in \mathcal{K}_{\text{exp}},\ \text{for each } \ i$$

Example 2 (Logistic regression)**.** Consider the simplest logistic regression

$$\max_{\beta} \sum_{i=1}^{n} y_i\left(x_i'\beta\right) - \log\left(1 + \exp\left(x_i'\beta\right)\right). \tag{16}$$

Introducing t_i to replace the softplus function in the objective and $\phi_i = x_i'\beta$, we turn (16) into

$$\max_{t_i,\phi_i,\beta} \sum_{i=1}^{n}\left(y_i\phi_i + t_i\right) \quad \text{s.t.} -\log\left(1+\exp\left(\phi_i\right)\right) \geq t_i,\ \ \phi_i = x_i'\beta \ \text{ for each } \ i.$$

Notice that $-\log\left(1+\exp\left(\phi_i\right)\right) \geq t_i$ is equivalent to $\exp\left(\phi_i + t_i\right) + \exp\left(t_i\right) \leq 1$. Introducing u_i and

v_i, we transform it into

$$u_i + v_i \leq 1, \quad (u_i, 1, \phi_i + t_i) \in \mathcal{K}_{\exp}, \quad (v_i, 1, t_i) \in \mathcal{K}_{\exp},$$

and then we reach the standard form

$$\begin{aligned} & \max_{\theta} \begin{bmatrix} \mathbf{1}_n & y & \mathbf{0}_{1\times(2n+p)} \end{bmatrix} \theta \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{0}_{n\times n} & \mathbf{I}_n & \mathbf{0}_{n\times 2n} & -X \\ \mathbf{0}_{n\times 2n} & \mathbf{I}_n & \mathbf{I}_n & \mathbf{0}_{n\times p} \end{bmatrix} \theta \begin{matrix} = \\ \leq \end{matrix} \begin{bmatrix} \mathbf{0}_{n\times 1} \\ \mathbf{1}_n \end{bmatrix} \\ & (u_i, 1, \phi_i + t_i) \in \mathcal{K}_{\exp}, \quad (v_i, 1, t_i) \in \mathcal{K}_{\exp}, \text{ for each } i \end{aligned}$$

where $\theta = (t, \phi, u, v, \beta)$.

- More details see Koenker and Mizera (2014) and Gao and Shi (2021).
- More code snippets, see https://github.com/zhan-gao/convex_prog_in_econometrics

4 Optimization-Conscious Econometrics

Textbook

- Course offered by Prof. Guillaume Allaire Pouliot

Examples

- Linear Programming
  - Candes and Tao (2007); Shi (2016) Dantzig Selector and L_∞-norm relaxation
  - Bright et al. (2025) matching
  - Koenker and Bassett Jr (1978); Koenker (2005) Quantile Regression
  - Carlier et al. (2016) Vector quantile regression via optimal transport
- Quadratic Programming
  - Shi et al. (2022) ℓ_2-relaxation relaxation in forecast combination
  - Hsieh et al. (2022)
- Conic Programming
  - Koenker and Mizera (2014); Gao and Shi (2021) Lasso variants.

  - Su et al. (2016) Classifier-Lasso

- Polynomial Programming
  - Lee (2022) Short-T dynamic panel

- Semidefinite Programming
  - Auerbach (2022) Approximation of cut-norm

- Mixed Integer Programming
  - Bertsimas et al. (2016) Best subset selection
  - Chen and Lee (2018) Max-score estimation
  - Kitagawa and Tetenov (2018) Optimal treatment rule
  - Pouliot (2023) Instrumental variable quantile regression (IVQR)

A Convex Sets and Convex Functions

A.1 Convex Set

- convex cone: set contains all conic conbinations of points in the set.
  - conic combination of x_1 and x_2 is $x = \theta_1 x_1 + \theta_2 x_2 \forall \theta_1, \theta_2 \geq 0$
- norm cone: $\{(x,t) \mid \|x\| \leq t\}$
- **Convexity preserving operations** - establish convexity by operations that preserves convexity from simple convex sets
  - **Intersection**
  - **Affine function**: image and inverse image of convex sets under $f(x) = Ax + b$ are convex
    - $\{x | x_1 A_1 + x_2 A_2 + \cdots + x_n A_n \preceq B\}$ with $A_i, B \in \mathrm{S}^p$
  - **Perspective function**: image and inverse image of convex sets under $P : \mathbb{R}^{n+1} \to \mathbb{R}^n, P(x,t) = \frac{x}{t}$ are convex
  - **Linear-fractional function**: image and inverse image of convex sets under $f : \mathbb{R}^n \to \mathbb{R}^m, f(x) = \frac{Ax+b}{c^\top x + d}$ are convex

A.2 Convex Function

- Examples:
  - $x \log x$ on $\mathbb{R}_{++}$
  - $f(X) = tr\left(A^\top X\right) + b = \sum_{i=1}^n \sum_{j=1}^m A_{ij} X_{ij} + b$
  - $f(X) = \|X\|_2 = \sigma_{\max}(X) = \left(\lambda_{\max}\left(X^\top X\right)\right)^{\frac{1}{2}}$
- **Restriction to a line**: f convex if and only if $g : \mathbb{R} \to \mathbb{R}$

$$g(t) = f(x + t\nu), \quad dom(g) = \{t | x + t\nu \in dom(f)\}$$

  is convex in t for any $x \in dom(f), \nu \in \mathbb{R}^n$

  - $f(X) = -\log \det X$ where $dom(f) = \mathrm{S}^n_{++}$
    - make use of $X^{\frac{1}{2}} \left(I - tX^{-\frac{1}{2}} V X^{-\frac{1}{2}}\right) X^{\frac{1}{2}}$

- **Fisrt order condition**: $f\left(y\right) \geq f\left(x\right) + \nabla f\left(x\right)^{\top}\left(y-x\right) \forall x, y \in dom\left(f\right)$

- **Second order condtion**: $\nabla^2 f\left(x\right) \succeq 0 \forall x \in dom\left(f\right)$
  - $f\left(x\right) = \log \sum_{k=1}^{n} \exp x_k$
  - $f\left(x\right) = \left(\prod_{k=1}^{n} x_k\right)^{1/n}$ on $\mathbb{R}_{++}^{n}$

- **Epigraph** convex iff f convex

- f convex $\Rightarrow$**sublevel sets** convex

Convexity preserving operations

- **Perspective**: perspective of f is $g\left(x,t\right) = tf\left(\frac{x}{t}\right)$. g is convex if f is convex.

- **Conjugate**: The conjugate of a function f: $f^*\left(y\right) = \sup_x \left\{y^{\top}x - f\left(x\right)\right\}$ is always convex even if f is not.

- **Pointwise maximum**: $f\left(x\right) = \max\left\{f_1\left(x\right), f_2\left(x\right), \cdots, f_m\left(x\right)\right\}$ is convex if f_j is convex for $j = 1, 2, \cdots, m$.

- **Composition with affine function**: $f\left(Ax+b\right)$ is convex if f is convex.
  - log barrier $\sum \log\left(b_i - a_i^{\top}x\right)$
  - norm of affine function $\|Ax+b\|$

- **Pointwise supreme**: $f\left(x,y\right)$ is convex in x for each $y \in \mathcal{A}$, then $g\left(x\right) = \sup_{y\in\mathcal{A}} f\left(x,y\right)$ is convex.
  - support function $S_C\left(x\right) = \sup_{x\in C} y^{\top}x$ is convex
  - $f\left(x\right) = \sup_{y\in c} \|x-y\|$
  - $\lambda_{\max}\left(X\right) = \sup_{\|y\|_2=1} y^{\top}Xy$

- **Minimization**: if $f\left(x,y\right)$ is convex in $\left(x,y\right)$ and C is a convex set, then $g\left(x\right) = \inf_{y\in C} f\left(x,y\right)$ is convex.

Remark 3. In pointwise supreme, the condition does not include the convexity of C and f is required to be convex in only x for each $y \in C$.

- **Composition with scalar functions**: $f\left(x\right) = h\left(g\left(x\right)\right)$
  f is convex if: g convex, h convex, h nondecreasing or g concave, h convex, h nonincreasing

- **Vector composition**: $f(x) = h(g(x)) = h(g_1(x), \cdots, g_k(x))$
  f is convex if: g_i convex, h convex, h nondecreasing in each argument or g_i concave, h convex, h nonincreasing in each argument
  - $\sum_{i=1}^{m} \log g_i(x)$ is concave
  - $\log \sum_{i=1}^{m} \exp g_i(x)$ is convex

A.3 Subgradient

- Subgradients defines a non-vertical supporting hyperplane to the epigraph *epi* (f) at $(x, f(x))$.
- $\partial f(x)$ is closed. If $x \in \mathrm{int}\,dom(f)$, then $\partial f(x)$ is nonempty and bounded.
- Monotonicity: $(u - v)^\top (x - y) \geq 0 \ \forall x, y, u \in \partial f(x), v \in \partial f(y)$

The followings are basic subdifferential rules and examples:

- $\partial |x| = \begin{cases} 1 & x > 0 \\ [-1, 1] & x = 0 \\ -1 & x < 0 \end{cases}$
- $\partial \|x\|_2 = \begin{cases} \frac{x}{\|x\|_2} & x \neq 0 \\ \{g | \, \|g\|_2 \leq 1\} & x = 0 \end{cases}$
- $h(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ with $\alpha_i \geq 0$, then $\partial h(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x)$
- $h(x) = f(Ax + b)$, then $\partial h(x) = A^\top \partial f(Ax + b)$
- **Pointwise Maximum**:Choose any active k and any subgradient f_k at x.
  - l_1-norm: $f(x) = \|x\|_1 = \max_{s \in \{-1,1\}^n} s^\top x$
- **Pointwise Supremum**: $f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$ where f_α is convex in x for any α. Find any active β and choose $g \in \partial f_\beta(x)$
  - $f(x) = \lambda_{\max}(A(x)) = \sup_{\|y\|_2 = 1} y^\top A(x) y$
- **Minimization**: $f(x) = \inf_y h(x, y)$, h jointly convex in (x, y). Find $\hat{y}$ that minimizes $h(\hat{x}, y)$ and find subgradient $(g, 0) \in \partial h(\hat{x}, \hat{y})$.

- $f(x) = \inf_{y \in C} \|x - y\|_2$ where C is a closed convex set.

  $\partial f(\hat{x}) = \begin{cases} 0 & \hat{x} \in C \\ \frac{\hat{x} - P(\hat{x})}{\|\hat{x} - P(\hat{x})\|_2} & \hat{x} \notin C \end{cases}$

- **Composition**: $f(x) = h(f_1(x), \cdots, f_k(x))$, h convex, non-increasing, f_i convex find $z \in \partial h(\hat{x})$ and $g_i \in \partial f_i(\hat{x})$, then $g = \sum_i z_i g_i \in \partial f(\hat{x})$

- **Optimal Value function**: if $\hat{\lambda}, \hat{\nu}$ are optimal dual variables for corresponding $\hat{u}, \hat{v}$, then $\left(-\hat{\lambda}, -\hat{\nu}\right) \in \partial h(\hat{u}, \hat{v})$.

- **Expectation**: $f(x) = \mathbb{E} h(x, u)$ where u random, h convex in x for every u choose a function $u \to g(u)$ with $g(u) \in \partial_x h(\hat{x}, u)$, then $g = \mathbb{E}_u g(u) \in \partial f(\hat{x})$

A.4 Proximal Gradient Method

$$\begin{aligned} x^k &= \arg\min_x \frac{1}{2t_k} \left\| x - \left(x^{k-1} - t_k \nabla g\left(x^{k-1}\right)\right)\right\|^2 + h(x) \\ &= \operatorname{prox}_{t_k h}\left(x^{k-1} - t_k \nabla g\left(x^{k-1}\right)\right) \end{aligned}$$

Through Moreau decomposition, proximal mapping is connected with conjugate functions.

A.4.1 Conjugate function

- **Fenchel's inequality**: $f(x) + f^*(y) \geq x^\top y \; \forall x, y$

- If f is closed and convex, then $f^{**}(x) = f(x) \forall x$

- If f is closed and convex, then $y \in \partial f(x) \Leftrightarrow x \in \partial f^*(y) \Leftrightarrow x^\top y = f(x) + f^*(y)$.

- **Examples**:
  - $f(X) = -\log\det X$, $f^*(Y) = -\log\det(-Y) - n$
    - $\nabla \langle X, Y \rangle = Y$, $\nabla \log\det X = X^{-1}$, $\left\langle Y, (-Y)^{-1} \right\rangle = -n$
  - Conjugate of indicator function of a convex set C is support function of C, $f^*(y) = \sup_{x \in C} y^\top x$
  - Conjugate of norm function is indicator of unit dual norm ball

Proximal mapping

- If h is closed and convex, then $\mathrm{prox}_h\left(x\right)$ exists (closed and bounded sublevel sets) and is unique (strong convexity)

- **Subgradient characterization**:

$$u = \mathrm{prox}_h\left(x\right) \Leftrightarrow x - u \in \partial h\left(u\right)$$

- **Nonexpansiveness (Lipschitz continuous with $L = 1$)** : if $u = \mathrm{prox}_h\left(x\right)$, $v = \mathrm{prox}_h\left(y\right)$, then

$$\left(u - v\right)^\top \left(x - y\right) \geq \left\|u - v\right\|_2^2 \Rightarrow \left\|x - y\right\|_2 \geq \left\|u - v\right\|_2$$

- **Moreau decomposition**:

$$x = \mathrm{prox}_f\left(x\right) + \mathrm{prox}_{f^*}\left(x\right) \forall x$$

- **Extended Moreau decomposition**: for $\lambda > 0$,

$$x = \mathrm{prox}_{\lambda f}\left(x\right) + \lambda \mathrm{prox}_{\lambda^{-1} f^*}\left(\frac{x}{\lambda}\right)$$

Examples

- $h\left(x\right) = I_C(x)$, then $\mathrm{prox}_h\left(x\right) = P_C\left(x\right)$

- $h\left(x\right) = \left\|x\right\|_1$, then $\mathrm{prox}_h\left(x\right)_i = \begin{cases} x_i - 1 & x_i \geq 1 \\ 0 & \left|x_i\right| \leq 1 \\ x_i + 1 & x_i \leq 1 \end{cases}$

- $f\left(x\right) = \left\|x\right\|_2$, then $\mathrm{prox}_{tf}\left(x\right) = \begin{cases} \left(1 - \frac{t}{\left\|x\right\|_2}\right) x & \left\|x\right\|_2 \geq t \\ 0 & \text{otherwise} \end{cases}$

Calculus rules

- Separable Sum: $f\left(\left(\begin{array}{c} x \\ y \end{array}\right)\right) = g\left(x\right) + h\left(y\right)$, $\mathrm{prox}_f\left(\left(\begin{array}{c} x \\ y \end{array}\right)\right) = \left(\begin{array}{c} \mathrm{prox}_{\mathrm{g}}\left(\mathrm{x}\right) \\ \mathrm{prox}_h\left(y\right) \end{array}\right)$

- Scaling and translation of argument with $\lambda \neq 0$: $f\left(x\right) = g\left(\lambda x + a\right)$, $\mathrm{prox}_f\left(x\right) = \frac{1}{\lambda}\left(\mathrm{prox}_{\lambda^2 g}\left(\lambda x + a\right) - a\right)$

- Scalar multiplication with $\lambda > 0$: $f\left(x\right) = \lambda g\left(\frac{x}{\lambda}\right)$, $\mathrm{prox}_f\left(x\right) = \lambda \mathrm{prox}_{\lambda^{-1} g}\left(\frac{x}{\lambda}\right)$

- Linear funtion: $f\left(x\right) = g\left(x\right) + a^\top x$, $\mathrm{prox}_f\left(x\right) = \mathrm{prox}_g\left(x - a\right)$

- Quadratic function with $\mu > 0$: $f(x) = g(x) + \frac{\mu}{2}\|x-a\|_2^2$, $\mathrm{prox}_f(x) = \mathrm{prox}_{\theta g}(\theta x + (1-\theta)a)$ where $\theta = \frac{1}{1+\mu}$
- Composition with affine mapping: $f(x) = g(Ax+b)$ where $AA^\top = \left(\frac{1}{\alpha}\right)I$, then

$$\mathrm{prox}_f(x) = \left(I - \alpha A^\top A\right)x + \alpha A^\top\left(\mathrm{prox}_{\alpha^{-1}g}(Ax+b) - b\right)$$

References

Auerbach, E. (2022). Testing for differences in stochastic network structure. *Econometrica 90*(3), 1205–1223.

Bertsekas, D. (1999). *Nonlinear programming* (2nd ed.). Athena Scientific.

Bertsimas, D., A. King, and R. Mazumder (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics 44*(2), 813–852.

Boyd, S., N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning 3*(1), 1–122.

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Bright, I., A. Delarue, and I. Lobel (2025). Reducing marketplace interference bias via shadow prices. *Management Science 71*(8), 7094–7112.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313 – 2351.

Carlier, G., V. Chernozhukov, and A. Galichon (2016). Vector quantile regression: An optimal transport approach. *The Annals of Statistics 44*(3), 1165 – 1192.

Chen, L.-Y. and S. Lee (2018). Best subset binary prediction. *Journal of Econometrics 206*(1), 39–56.

Gao, Z. and Z. Shi (2021). Implementing convex optimization in R: Two econometric examples. *Computational Economics 58*(4), 1127–1135.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Hsieh, Y.-W., X. Shi, and M. Shum (2022). Inference on estimators defined by mathematical programming. *Journal of Econometrics 226*(2), 248–268.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Koenker, R. (2005). *Quantile regression*, Volume 38. Cambridge university press.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica*, 33–50.

Koenker, R. and I. Mizera (2014). Convex optimization in R. *Journal of Statistical Software 60*(5), 1–23.

Lee, W. (2022). Identification and estimation of dynamic random coefficient models. Working paper.

Pouliot, G. A. (2023). Instrumental variables quantile regression with multivariate endogenous variable. *Unpublished Working Paper*.

Shi, Z. (2016). Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics 195*(1), 104–119.

Shi, Z., L. Su, and T. Xie (2022, 11). ℓ2-Relaxation: With Applications to Forecast Combination and Portfolio Analysis. *The Review of Economics and Statistics*, 1–44.

Su, L., Z. Shi, and P. C. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*(6), 2215–2264.