Statistics and Optimization
Convex optimization for thresholding estimators

Sylvain Sardy

Université de Genève

1. Subgradients and subdifferential

Remember that a differentiable function f is convex on a convex set $C \subset \mathbb{R}^N$ iff

$$f(\mathbf{z}) \geq f(\mathbf{x}) + (\mathbf{z} - \mathbf{x})^{\mathrm{T}} \nabla f(\mathbf{x}) \quad \forall \mathbf{x}, \mathbf{z} \in C.$$

Definition: a *subgradient* of a convex function f at $\mathbf{x}$ is any $\mathbf{g} \in \mathbb{R}^N$ such that

$$f(\mathbf{z}) \geq f(\mathbf{x}) + (\mathbf{z} - \mathbf{x})^{\mathrm{T}} \mathbf{g} \quad \forall \mathbf{z} \in C.$$

Definition: the *subdifferential* of a convex function f at $\mathbf{x}$ is the set of all subgradients

$$\partial f(\mathbf{x}) := \{\mathbf{g} \in \mathbb{R}^N : \ \mathbf{g} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}.$$

Geometric interpretation: $\mathbf{g}$ indexes hyperplanes tangent to the graph of f at $\mathbf{x}$.

Examples:

- $f(x) = |x|$. The subdifferential of f at x is

$$\partial f(x) := \begin{cases} \{\mathrm{sign}(x)\} & \text{if } x \neq 0 \\ [-1, 1] & \text{if } x = 0 \end{cases} .$$

- $f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{n=1}^{N} f_n(\mathbf{x})$ with $f_n(\mathbf{x}) = |x_n|$. The subdifferential of f at x is

$$\partial f(\mathbf{x}) := \sum_{n=1}^{N} s_n \mathbf{e}_n \quad \text{with} \quad s_n = \begin{cases} \{\mathrm{sign}(x_n)\} & \text{if } x_n \neq 0 \\ [-1, 1] & \text{if } x_n = 0 \end{cases} ,$$

  where $\mathbf{e}_n$ is the nth canonical vector.

- $f(\mathbf{x}) = \|\mathbf{x}\|_2$. The subdifferential of f at x is

$$\partial f(\mathbf{x}) := \begin{cases} \{\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\} & \text{if } \mathbf{x} \neq \mathbf{0} \\ \{\mathbf{z} : \ \|\mathbf{z}\|_2 \leq 1\} & \text{if } \mathbf{x} = \mathbf{0} \end{cases} .$$

2. Calculus of subdifferential

Properties:

- f is differentiable at $\mathbf{x}$ iff $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.
- $\partial f(\mathbf{x})$ is closed and convex, and nonempty.
- $\partial\left(\lambda f(\mathbf{x})\right) = \lambda \partial f(\mathbf{x})$ if $\lambda \geq 0$.
- $\partial\left(f_1(\mathbf{x}) + f_2(\mathbf{x})\right) = \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x})$.
- $\partial f(A\mathbf{x} + \mathbf{b}) = A^{\mathrm{T}} \partial f(A\mathbf{x} + \mathbf{b})$.

Optimality condition

For any convex function $f : \mathbb{R}^N \to \mathbb{R}$, a point $\mathbf{x}^* \in \mathbb{R}^N$ is a global minimum of f iff the condition $\mathbf{0} \in \partial f(\mathbf{x}^*)$ holds:

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^N} \Leftrightarrow \mathbf{0} \in \partial f(\mathbf{x}^*).$$

3. Some simple applications in Statistics

- Example 1: $\min_w \frac{1}{2}(y-w)^2 + \lambda|w|$.
  A minimum exists and is unique, and

$$\partial f(w) = \begin{cases} \{w - y - \lambda\} & \text{if } w < 0, \\ [w - y - \lambda, w - y + \lambda] & \text{if } w = 0, \\ \{w - y + \lambda\} & \text{if } w > 0. \end{cases}$$

So

$$0 \in \partial f(w^*) \quad \Leftrightarrow \quad \begin{cases} w^* = y + \lambda & \text{if } y < -\lambda \\ w^* = 0 & \text{if } y \in [-\lambda, \lambda] \\ w^* = y - \lambda & \text{if } y > \lambda \end{cases}$$

$$\Leftrightarrow \quad w^* = \max[(1 - \frac{\lambda}{|y|}), 0]y.$$

The function $\eta_\lambda^{\text{soft}}(y) = \max[(1 - \frac{\lambda}{|y|}), 0]y$
is the *soft-thresholding* function
(Donoho and Johnstone, 1994).

- Example 2: solve

$$\min_{\alpha \in \mathbb{R}} \frac{1}{2} \|\mathbf{y} - \mathbf{x}\alpha\|_2^2 + \lambda |\alpha|,$$

where $\mathbf{y}$ and $\mathbf{x}$ are in $\mathbb{R}^N$.

- Example 3: solve

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \frac{1}{2} \|\mathbf{y} - \Phi\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1,$$

where $\mathbf{y} \in \mathbb{R}^N$ and Φ is is an orthonormal matrix.

This estimation problem occurs for denoising a signal using wavelets.

What is a signal denoising problem?
A univariate function $\mu(x)$ observed with noise

$$Y_n = \mu(x_n) + \epsilon_n \quad n = 1, \ldots, N.$$

The goal is to recover μ from noisy observations $\mathbf{Y}$.

One approach assumes

$$\mu(x) = \sum_j \sum_k \phi_{j,k}(x) \alpha_{j,k},$$

where $\phi_{j,k}$ are known orthonormal basis functions indexed by level j and translation k.

Haar wavelets (1910):

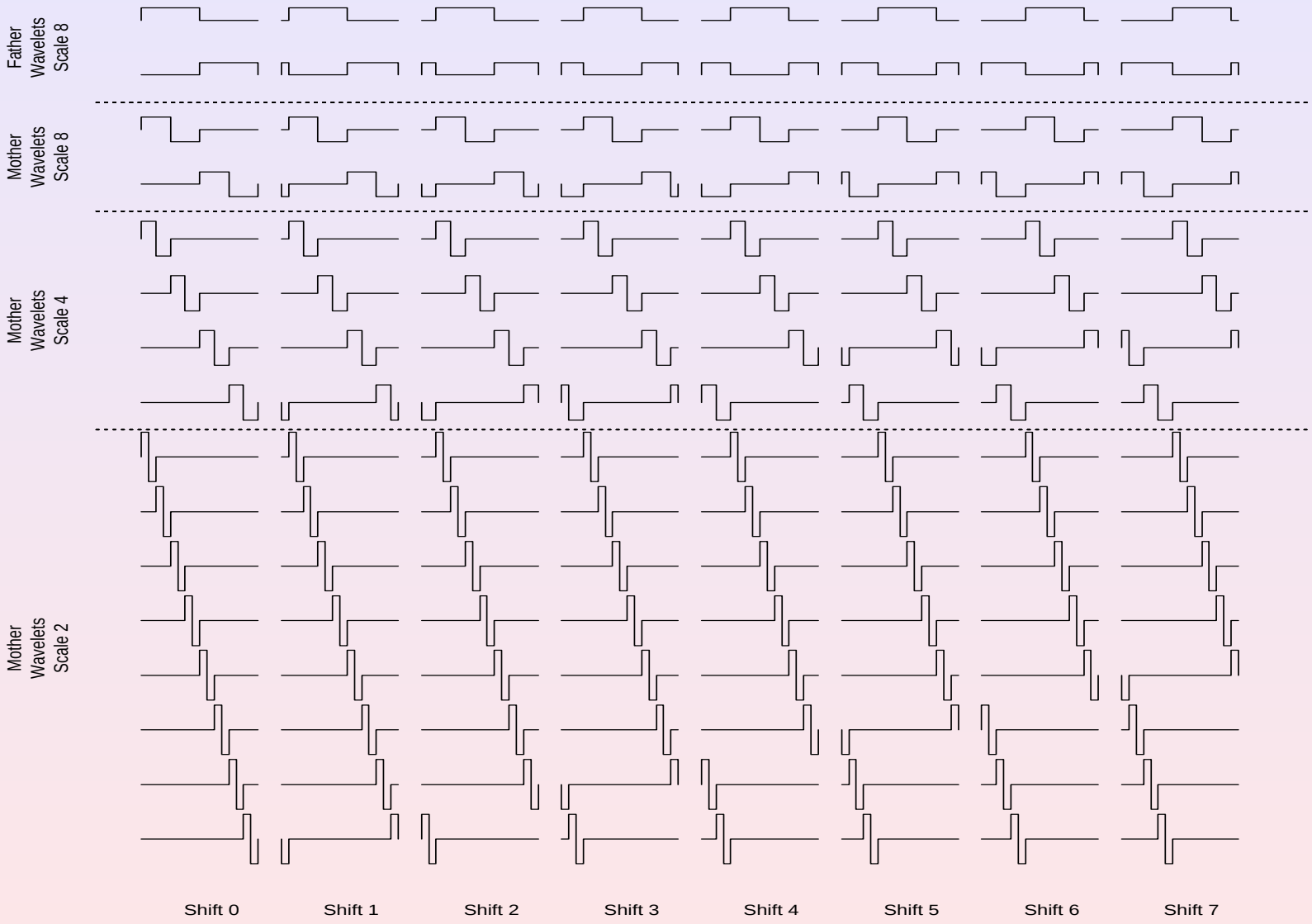
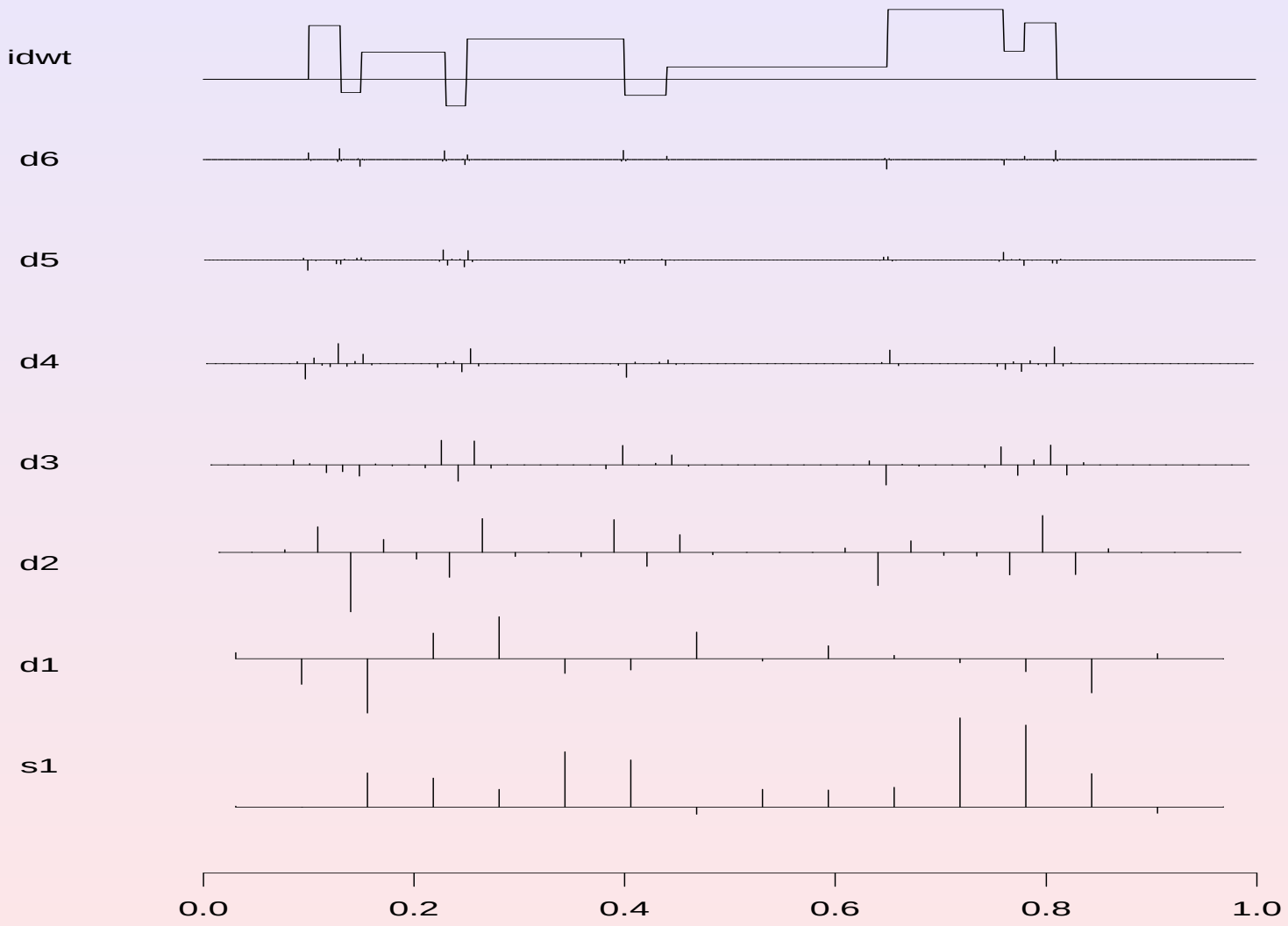
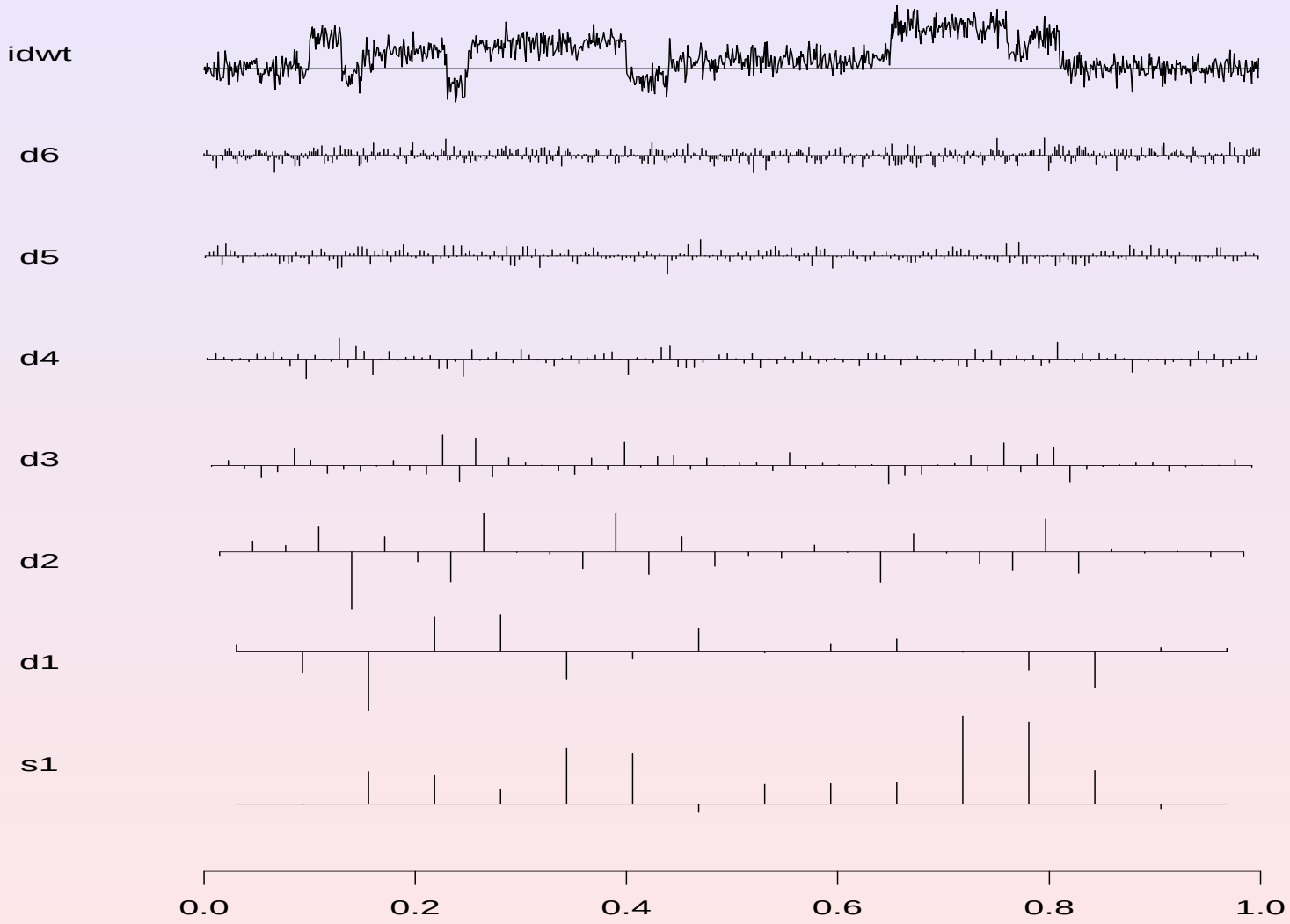


Illustration of sparse wavelet representation:



A noisy signal and its wavelet representation:



Once discretized and vectorized the wavelet-based model becomes

$$\mathbf{Y} = \Phi\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

To achieve sparse estimation of $\boldsymbol{\alpha}$, Donoho and Johnstone (1994) define the soft-waveshrink estimator which soft-thresholds

$$\hat{\boldsymbol{\alpha}}_\lambda = \eta_\lambda^{\mathrm{soft}}(\hat{\boldsymbol{\alpha}}^{\mathrm{LS}})$$

the least squares wavelet coefficients

$$\hat{\boldsymbol{\alpha}}^{\mathrm{LS}} = \Phi^{\mathrm{T}}\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\alpha}, \sigma^2 I_N).$$

Interestingly

$$\eta_\lambda^{\mathrm{soft}}(\hat{\boldsymbol{\alpha}}^{\mathrm{LS}}) = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \Phi\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1.$$

Introduction to duality problem

Given the (primal) general problem

$$\min_{\mathbf{x}\in C} f(\mathbf{x}) \quad \text{with} \quad C = \{\mathbf{x} \in \mathbb{R}^N : \begin{cases} h_i(\mathbf{x}) \leq 0 & i = 1, \ldots, m \\ l_j(\mathbf{x}) = 0 & j = 1, \ldots, r \end{cases} \},$$

define the Lagrangian as

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \sum_i u_i h_i(\mathbf{x}) + \sum_j v_j l_j(\mathbf{x})$$

with $\mathbf{u} \geq 0$ and $L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = -\infty$ for $\mathbf{u} < \mathbf{0}$.

Clearly $f(\mathbf{x}) \geq L(\mathbf{x}, \mathbf{u}, \mathbf{v})$ for all $\mathbf{x} \in C$.
so $f^* \geq \min_{\mathbf{x}\in C} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \geq \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) =: g(\mathbf{u}, \mathbf{v})$.

Definition: $g(\mathbf{u}, \mathbf{v})$ is the Lagrange dual function and $(\mathbf{u}, \mathbf{v})$ are the dual variables.

Clearly the best lower bound of f^* based on the dual function is

$$g^* = \max_{\mathbf{u} \geq \mathbf{0}, \mathbf{v}} g(\mathbf{u}, \mathbf{v}) \quad \text{(Dual problem)}.$$

Consequently $f^* \geq g^*$, called *weak duality*.

Strong duality $f^* = g^*$ holds assuming Slater's condition: f and h_i are convex (convex program) and if there exists $\mathbf{x}$ strictly satisfying the non-affine inequality contraints.

What is the use of duality?

- maximizing the dual problem (which may be easier to solve) leads to a solution to the primal problem.
- the duality gap

$$\delta = f(\mathbf{x}) - g(\mathbf{u}, \mathbf{v}) \geq 0$$

should be zero at $(\mathbf{x}, \mathbf{u}, \mathbf{v}) = (\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*)$ when strong duality holds (can be used to check convergence of algorithm).

Karush-Kuhn-Tucker conditions

Given the (primal) general problem

$$\min_{\mathbf{x}\in C} f(\mathbf{x}) \quad \text{with} \quad C = \left\{ \mathbf{x} \in \mathbb{R}^N : \left\{ \begin{array}{ll} h_i(\mathbf{x}) \leq 0 & i = 1, \ldots, m \\ l_j(\mathbf{x}) = 0 & j = 1, \ldots, r \end{array} \right. \right\},$$

the first order KKT optimality conditions are

$$\left\{ \begin{array}{l} \mathbf{0} \in \partial f(\mathbf{x}) + \sum_i u_i \partial h_i(\mathbf{x}) + \sum_j v_j \partial l_j(\mathbf{x}) \\ u_i h_i(\mathbf{x}) = 0 \ \forall i \\ h_i(\mathbf{x}) \leq 0, \ l_j(\mathbf{x}) = 0 \ \forall i, j \\ u_i \geq 0 \ \forall i \end{array} \right. .$$

Assuming strong duality holds,

$$(\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*) \text{ are primal and dual solutions}$$

$$\Updownarrow$$

$$(\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*) \text{ satisfy the KKT conditions.}$$