OPTIMIZATION THEORY

Reference:

J. Nocedal and S.J. Wright, "*Numerical Optimization*," 2006. Chapter 2

OPTIMALITY CONDITIONS

THEOREM (TAYLOR'S THEOREM)

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable and* $p \in \mathbb{R}^n$*. Then for some* $t \in (0,1)$ *we have that*

$$f(x+p) = f(x) + \nabla f(x+tp)'p$$

Moreover, if f *is twice continuously differentiable, for some* $t \in (0,1)$ *we have that*

$$f(x+p) \quad = \quad f(x) + \nabla f(x)'p + \frac{1}{2}p'\nabla^2 f(x+tp)p$$

THEOREM (FIRST-ORDER NECESSARY CONDITIONS)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and x^* a local optimizer. Then

$$\nabla f(x^*) = 0$$

Proof:

- Assume by contradiction that $p = -\nabla f(x^*) \neq 0$. Let $g(t) = p'\nabla f(x^* + tp)$. Then $g(0) = p'\nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0$
- ∇f is continuous around x^*, so g is also continuous wrt t in $t = 0$, and therefore $\exists T > 0$ such that $g(t) < 0$ for all $t \in [0, T]$
- For any $\bar{t} \in (0, T]$ by Taylor's theorem we have that for some $t \in (0, \bar{t})$

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p'\nabla f(x^* + tp) = f(x^*) + g(t)\bar{t} < f(x^*),\ \forall \bar{t} \in (0, T]$$

- Then x^* is not a local minimizer, which is a contradiction. $\square$

THEOREM (SECOND-ORDER NECESSARY CONDITIONS)

*Let the **Hessian** matrix function $\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ exist and be continuous in an open neighborhood of a local optimizer x^*. Then*

$$\nabla f(x^*) = 0, \nabla f^2(x^*) \succeq 0$$

Proof:

- Assume by contradiction that $\nabla^2 f(x^*) \not\succeq 0$. Then there exist p such that $p'\nabla^2 f(x^*)p < 0$.
- Since $\nabla^2 f(x)$ is continuous around x^*, $\exists T > 0$ such that $p'\nabla^2 f(x^* + tp)p < 0$ for all $t \in [0, T]$.
- By doing a Taylor expansion around x^*, $\forall \bar{t} \in (0, T]$ there exists $t \in (0, \bar{t})$ such that

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p'\nabla f(x^*) + \frac{1}{2}\bar{t}^2 p'\nabla^2 f(x^* + tp)p < f(x^*)$$

- Then x^* is not a local minimizer, which is a contradiction. $\square$

THEOREM (SECOND-ORDER SUFFICIENT CONDITIONS)

Let $\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ exist and be continuous in an open neighborhood of x^*. Let $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$. Then x^* is a strict local minimizer of f.

Proof:

- Since the Hessian function $\nabla^2 f(x)$ is continuous in x^* and $\nabla^2 f(x^*) \succ 0$, $\nabla^2 f(x) \succ 0$ for all x in an open ball $B(x^*, r)$[1] for some scalar $r > 0$
- For any p such that $\|p\|_2 < r$ we have that $x^* + p \in B(x^*, r)$ and hence

$$f(x^*+p) = f(x^*)+p'\nabla f(x^*)+\frac{1}{2}p'\nabla^2 f(x^*+tp)p = f(x^*)+\frac{1}{2}p'\nabla^2 f(x^*+tp)p$$

for some $t \in (0, 1)$.

- Since $x^* + tp \in B(x^*, r)$, $p'\nabla^2 f(x^* + tp)p > 0$, and therefore $f(x^* + p) > f(x^*)$, $\forall p \in B(0, r)$. □

[1] For a positive scalar $r > 0$, the **Euclidean ball** $B(x_0, r)$ is the set $\{x : \|x - x_0\|_2 \le r\}$.

OPTIMALITY CONDITIONS - CONSTRAINED CASE

- Consider the constrained optimization problem

$$\begin{array}{rl} \min_x & f(x) \\ \text{s.t.} & g_i(x) \leq 0,\ i \in I \\ & g_j(x) = 0,\ j \in E \end{array}$$

  with $I \cup E = \{1, \ldots, m\}$.

- A vector x is **feasible** if $g_i(x) \leq 0, \forall i \in I$, and $g_j(x) = 0, \forall j \in E$

- We say that the inequality constraint $i \in I$ is **active** if $g_i(x) = 0$, **inactive** if $g_i(x) < 0$ (equality constraints $g_j(x)$, $j \in E$, are always active).

OPTIMALITY CONDITIONS - CONSTRAINED CASE

- The **active set** $\mathcal{A}(x)$ at any feasible vector x is the set of indexes

$$\mathcal{A}(x) = \{i \in I : g_i(x) = 0\} \cup E$$

- We say that the **linear independence constraint qualification** (LICQ) condition holds at x if the vectors $\{\nabla g_i(x)\}_{i \in \mathcal{A}(x)}$ are linearly independent

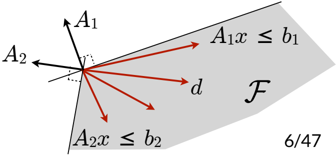
- The set $\mathcal{F}(x)$ of **linearized feasible directions** at a feasible x is the cone

$$\mathcal{F}(x) = \{d : \ d'\nabla g_i(x) = 0, \ \forall i \in E, \ d'\nabla g_i(x) \le 0, \ \forall i \in \mathcal{A}(x), \ i \notin E\}$$

  Note that $g_i(x+d) \approx \underbrace{g_i(x)}_{=0} + \nabla g_i(x)'d$ for $d \to 0, \forall i \in \mathcal{A}(x)$

- Linear case example:

$$\left\{ \begin{array}{rcl} A_1 x & \le & b_1 \\ A_2 x & \le & b_2 \end{array} \right. \quad \Rightarrow \quad \left\{ \begin{array}{rcl} A_1 d & \le & 0 \\ A_2 d & \le & 0 \end{array} \right.$$



THEOREM

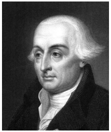
If x^* is a local minimum and the LICQ condition is satisfied then

$$\nabla f(x^*)'d \geq 0,\ \forall d \in \mathcal{F}(x^*)$$

- Define the **Lagrangian function**

$$\mathcal{L}(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

where $\lambda \in \mathbb{R}^m$ are the **Lagrange multipliers**,
$I \cup E = \{1,\ldots,m\}$



Joseph-Louis Lagrange
(1736–1813)

KKT OPTIMALITY CONDITIONS

THEOREM (FIRST-ORDER NECESSARY CONDITIONS)

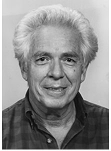
Let f and $g_i, i = 1, \ldots, m$, be continuously differentiable and x^* a local optimizer. Let the LICQ condition hold at x^*. Then $\exists \lambda^* \in \mathbb{R}^m$ such that

Karush Kuhn Tucker (KKT) conditions

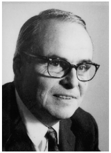
$$
\begin{array}{rcll}
\nabla_x \mathcal{L}(x^*, \lambda^*) & = & 0 & \\
g_i(x^*) & \leq & 0 & \forall i \in I \\
g_i(x^*) & = & 0 & \forall i \in E \\
\lambda_i^* & \geq & 0 & \forall i \in I \\
\lambda_i^* g_i(x^*) & = & 0 & \forall i = 1, \ldots, m
\end{array}
$$

- $\lambda_i^* g_i(x^*) = 0$ is a **complementary slackness** condition
- **strict complementarity** holds if $\lambda_i^* > 0$ for all $i \in \mathcal{A}(x^*)$
- λ^* is unique if the LICQ condition holds

William Karush (1917–1997)

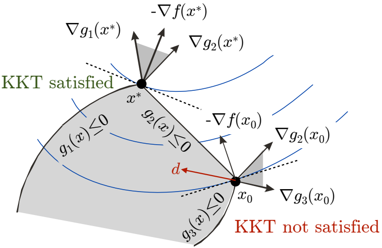


Harold W. Kuhn (1925–2014)



Albert W. Tucker (1905–1995)

KKT OPTIMALITY CONDITIONS



$$-\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*),\ \lambda_i^* \geq 0,\ E = \emptyset$$

$$f(x^* + \epsilon d) \approx f(x^*) + \epsilon \nabla f(x^*)'d$$

f decreases when $-\nabla f(x^*)'d > 0$

- if $-\nabla f(x^*)'d = \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*)'d$ were positive then $\nabla g_i(x^*)'d > 0$ for some $i \in \mathcal{A}(x^*)$ such that $\lambda_i^* > 0$.

  Hence f can only decrease at x^* if some active constraint g_i is violated, as $g_i(x^* + \epsilon d) \approx g_i(x^*) + \epsilon \nabla g_i(x^*)'d = \epsilon \nabla g_i(x^*)'d > 0$, $\epsilon > 0$

- Vice versa, if $-\nabla f(x^*)$ does not belong to the convex cone one can move in a direction d such that $d'\nabla f(x^*) < 0$ (that is, decrease f) while keeping $g_i(x) \leq 0$

KKT CONDITIONS FOR EQUALITY-CONSTRAINED QP

- **Quadratic programming** problem subject to **equality** constraints:

$$\begin{array}{ll} \min & \frac{1}{2}x'Qx + c'x \\ \text{s.t.} & Ax = b \end{array} \qquad Q = Q' \succ 0,\ A \text{ full row rank}$$

- Lagrangian function: $\mathcal{L}(x,\lambda) = \frac{1}{2}x'Qx + c'x + \lambda'(Ax - b)$

- KKT conditions:

$$\begin{array}{l} Qx + c + A'\lambda = 0 \\ Ax = b \end{array} \quad \Rightarrow \quad \begin{array}{l} x = -Q^{-1}(c + A'\lambda) \\ AQ^{-1}A'\lambda = -(b + AQ^{-1}c) \end{array}$$

  and therefore

$$\begin{aligned} \lambda^* &= -(AQ^{-1}A')^{-1}(b + AQ^{-1}c) \\ x^* &= -Q^{-1}(c - A'(AQ^{-1}A')^{-1}(b + AQ^{-1}c)) \end{aligned}$$

- In this case, the KKT conditions are also **sufficient** for optimality (this is a convex optimization problem, see later ...)

KKT CONDITIONS FOR QP

- **Quadratic programming** problem

$$\begin{array}{rl} \min & \frac{1}{2}x'Qx + c'x \\ \text{s.t.} & Ax \le b \\ & Ex = f \end{array}$$

- Lagrangian function: $\mathcal{L}(x,\lambda,\nu) = \frac{1}{2}x'Qx + c'x + \lambda'(Ax-b) + \nu'(Ex-f)$

- KKT conditions:

$$\begin{aligned} & Qx + c + A'\lambda + E'\nu = 0 \\ & Ex = f \\ & Ax \le b \\ & \lambda \ge 0 \\ & \lambda'(Ax-b) = 0 \end{aligned}$$

where we replaced $\lambda_i(A_i x - b_0) = 0, \forall i$, with $\sum_i \lambda_i(A_i x - b_0) = 0$, having imposed $\lambda_i \ge 0$, $A_i x \le b_i, \forall i$

2ND-ORDER NECESSARY CONDITIONS

SUPPLEMENTARY MATERIAL

- Let x^*, λ^* satisfy the KKT conditions. The **critical cone** $\mathcal{C}(x^*, \lambda^*)$ is defined as

$$\mathcal{C}(x^*, \lambda^*) = \left\{ w : \begin{array}{ll} \nabla g_i(x^*)'w = 0, & \forall i \in E \\ \nabla g_i(x^*)'w = 0, & \forall i \in \mathcal{A}(x^*) \cap I \text{ with } \lambda_i^* > 0 \\ \nabla g_i(x^*)'w \leq 0, & \forall i \in \mathcal{A}(x^*) \cap I \text{ with } \lambda_i^* = 0 \end{array} \right\}$$

- The critical cone $\mathcal{C}(x^*, \lambda^*)$ contains directions in $\mathcal{F}(x^*)$ for which it is not clear from gradient information only whether f will increase or decrease, as from the KKT conditions we have

$$w'\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i^* w' \nabla g_i(x^*) = 0, \ \forall w \in \mathcal{C}(x^*, \lambda^*)$$

2ND-ORDER CONDITIONS

SUPPLEMENTARY MATERIAL

THEOREM (2ND-ORDER NECESSARY CONDITIONS)

Assume f, g be twice continuously differentiable. Let x^* be a local minimum and the LICQ condition satisfied and λ^* such that the KKT conditions are satisfied. Then

$$w'\nabla_{xx}\mathcal{L}(x^*, \lambda^*)w \geq 0, \ \forall w \in \mathcal{C}(x^*, \lambda^*)$$

THEOREM (2ND-ORDER SUFFICIENT CONDITIONS)

Assume f, g be twice continuously differentiable. Let x^*, λ^* satisfy the KKT conditions and assume that

$$w'\nabla_{xx}\mathcal{L}(x^*, \lambda^*)w > 0, \ \forall w \in \mathcal{C}(x^*, \lambda^*), \ w \neq 0$$

Then x^* is a strict local minimum.

SENSITIVITY ANALYSIS

- Question: if we slightly perturb a constraint g_i how much $f(x^*)$ will change?
- The Lagrange multipliers λ^* answer such a **sensitivity analysis** question
- If $g_i(x^*) < 0$ ($\Rightarrow \lambda_i^* = 0$), perturbing $g_i(x^*) \leq 0$ to $g_i(x^*) \leq -\epsilon$ does not change the solution, $\forall \epsilon < -g_i(x^*)$, as the same x^*, λ^* satisfy the KKT
- Let us change one of the active constraints $g_i(x) \leq 0$ to $g_i(x) \leq -\epsilon$, $i \in \mathcal{A}(x^*)$
- Let $x^*(\epsilon)$ be the perturbed optimal solution and assume $|\epsilon|$ small enough so that $\mathcal{A}(x^*(\epsilon)) = \mathcal{A}(x^*)$

- By taking the Taylor expansion of $g_i(x^*(\epsilon))$ around $\epsilon = 0$ we get

$$g_j(x^*(\epsilon)) - g_j(x^*) \approx \nabla g_j(x^*)'(x^*(\epsilon) - x^*),\ j = 1, \ldots, m$$

- Since we assumed $\mathcal{A}(x^*(\epsilon)) = \mathcal{A}(x^*)$, then $g_i(x^*(\epsilon)) = -\epsilon$ and $g_j(x^*(\epsilon)) = 0$, $\forall j \in \mathcal{A}(x^*) \setminus \{i\}$, in addition to $g_j(x^*) = 0$, $\forall j \in \mathcal{A}(x^*)$

- By expanding $f(x^*(\epsilon))$ around $\epsilon = 0$ and using the KKT conditions

$$\begin{aligned} f(x^*(\epsilon)) - f(x^*) \approx \nabla f(x^*)'(x^*(\epsilon) - x^*) &= \sum_{j \in \mathcal{A}(x^*)} -\lambda_j^* \nabla g_j(x^*)'(x^*(\epsilon) - x^*) \\ &= \sum_{j \in \mathcal{A}(x^*)} -\lambda_j^*(g_j(x^*(\epsilon)) - g_j(x^*)) = \epsilon \lambda_i^* \end{aligned}$$

- For $\epsilon \to 0$ we get

$$\boxed{\frac{df(x^*(\epsilon))}{d\epsilon} = \lambda_i^*}$$

DEFINITION

Let $i \in \mathcal{A}(x^*)$. An inequality constraint g_i is **strongly active** if $\lambda_i^* > 0$, **weakly active** if $\lambda_i^* = 0$

- If a constraint is weakly active, modifying it slightly does not change the optimal value since $\frac{df(x^*(\epsilon))}{d\epsilon} = 0$
- Let us scale the constraints to $\beta_i g_i(x) \leq 0$, $\beta_i > 0$. The KKT conditions are satisfied for x^* and $\frac{\lambda_i^*}{\beta_i}$
- For the consistent perturbation of the constraint $\beta_i g_i(x) \leq -\beta_i \epsilon$ we get the same optimizer $x^*(\epsilon)$, and moreover the sensitivity at the solution is

$$\frac{\lambda_i^*}{\beta_i} = \frac{df(x^*(\epsilon))}{d(\beta_i \epsilon)} = \frac{1}{\beta_i} \frac{df(x^*(\epsilon))}{d\epsilon} \quad \Rightarrow \quad \frac{df(x^*(\epsilon))}{d\epsilon} = \lambda_i^*$$

DUALITY

- Consider again the optimization problem

$$\begin{array}{rl} \min_x & f(x) \\ \text{s.t.} & g_i(x) \leq 0,\ i \in I \\ & g_j(x) = 0,\ j \in E \end{array} \qquad I \cup E = \{1, \ldots, m\}$$

- Define the **dual function** $q : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$

$$\boxed{q(\lambda) = \inf_x \mathcal{L}(x, \lambda)} = \inf_x \left\{ f(x) + \sum_{i=1}^m \lambda_i g_i(x) \right\}$$

- The **domain** $\mathcal{D}$ of q is the set of all λ for which $q(\lambda) > -\infty$
- A vector $\lambda \in \mathcal{D}$ is **dual feasible** if $\lambda_i \geq 0, \forall i \in I$
- A vector is $x \in \mathbb{R}^n$ **primal feasible** if $g_i(x) \leq 0, \forall i \in I$ and $g_j(x) = 0, \forall j \in E$

THEOREM (WEAK DUALITY)

For any given ***primal feasible*** x *and* ***dual feasible*** λ

$$\boxed{q(\lambda) \leq f(x)}$$

In particular $q(\lambda) \leq f(x^*)$.

Proof:

- Since x and λ are feasible, $\lambda_i g_i(x) \leq 0, \forall i \in I$ and $\lambda_j g_j(x) = 0, \forall j \in E$
- Therefore

$$f(x) \geq f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) = \mathcal{L}(x, \lambda) \geq \inf_x \mathcal{L}(x, \lambda) = q(\lambda)$$

- Since the above relation holds for all feasible x, in particular it holds for x^*

$$f(x^*) \geq q(\lambda), \ \forall \lambda \text{ such that } \lambda_i \geq 0, \ i \in I$$

$\square$

THEOREM

*The dual function $q(\lambda)$ is **concave** and its domain $\mathcal{D}$ is **convex**.*

Proof:

- Take any $\lambda^1, \lambda^2 \in \mathcal{D}$, and $\alpha \in [0,1]$. We want to verify that $\alpha\lambda^1 + (1-\alpha)\lambda^2 \in \mathcal{D}$ and that Jensen's inequality holds:

$$
\begin{aligned}
q(\alpha\lambda^1 + (1-\alpha)\lambda^2) &= \inf_x \mathcal{L}(x, \alpha\lambda^1 + (1-\alpha)\lambda^2) \\
&= \inf_x \left\{ f(x) + \sum_{i=1}^m (\alpha\lambda_i^1 + (1-\alpha)\lambda_i^2) g_i(x) \right\} \\
&= \inf_x \left\{ (\alpha + 1 - \alpha) f(x) + \alpha \sum_{i=1}^m \lambda_i^1 g_i(x) + (1-\alpha) \sum_{i=1}^m \lambda_i^2 g_i(x) \right\} \\
&= \inf_x \left\{ \alpha \Big(f(x) + \sum_{i=1}^m \lambda_i^1 g_i(x)\Big) + (1-\alpha)\Big(f(x) + \sum_{i=1}^m \lambda_i^2 g_i(x)\Big) \right\} \\
&\geq \inf_{x_1} \left\{ \alpha \Big(f(x_1) + \sum_{i=1}^m \lambda_i^1 g_i(x_1)\Big) \right\} + \inf_{x_2} \left\{ (1-\alpha)\Big(f(x_2) + \sum_{i=1}^m \lambda_i^2 g_i(x_2)\Big) \right\}
\end{aligned}
$$

- Finally, we get

$$q(\alpha\lambda^1 + (1-\alpha)\lambda^2) \geq \alpha q(\lambda^1) + (1-\alpha)q(\lambda^2) > -\infty$$

which proves that q is concave and that $\alpha\lambda^1 + (1-\alpha)\lambda^2 \in \mathcal{D}$ $\square$

- Recall that the minimum of a finite number of affine functions is concave. $q(\lambda)$ is the minimum of infinitely many affine functions (one for each x).

DUAL PROBLEM

- We define **dual problem** of a given optimization problem the new problem

$$\begin{array}{rl} \max_{\lambda\in\mathbb{R}^n} & q(\lambda) \\ \text{s.t.} & \lambda_i \geq 0,\ \forall i \in I \end{array}$$

- The dual problem is always a convex programming problem, even if the primal problem is not convex

- Since $f(x^*) \geq q(\lambda)$ for all dual feasible λ, we also have that the optimum of the dual problem satisfies the **weak duality** condition

$$q(\lambda^*) \leq f(x^*)$$

- **Strong duality** holds when $q(\lambda^*) = f(x^*)$

- The difference $f(x^*) - q(\lambda^*)$ is called **duality gap**

GRADIENT OF DUAL FUNCTION AND ITS LINEAR APPROXIMATION

- Let $x^*(\lambda) = \arg\min_x \mathcal{L}(x,\lambda)$. For all $\lambda \geq 0$, the gradient

$$\boxed{\nabla_\lambda q(\lambda) = g(x^*(\lambda))}$$

Proof:

$$\nabla_\lambda q(\lambda) = \nabla_\lambda(\inf_x \mathcal{L}(x,\lambda)) = \nabla_\lambda \mathcal{L}(x^*(\lambda),\lambda)$$

$$= \nabla_\lambda x^*(\lambda) \underbrace{\frac{\partial \mathcal{L}(x^*(\lambda),\lambda)}{\partial x}}_{= 0 \text{ by optimality of } x^*(\lambda)} + \underbrace{\frac{\partial \mathcal{L}(x^*(\lambda),\lambda)}{\partial \lambda}}_{= g(x^*(\lambda))}$$

- The first-order Taylor expansion of the dual function around λ_0 is

$$\boxed{q(\lambda) \approx f(x^*(\lambda_0)) + g(x^*(\lambda_0))'\lambda}$$

- Proof:

$$\begin{aligned} q(\lambda) &\approx q(\lambda_0) + \nabla_\lambda q(\lambda_0)'(\lambda-\lambda_0) = q(\lambda_0) + g(x^*(\lambda_0))'(\lambda-\lambda_0) \\ &= \inf_x \mathcal{L}(x,\lambda_0) + g(x^*(\lambda_0))'(\lambda-\lambda_0) = f(x^*(\lambda_0)) + g(x^*(\lambda_0))'\lambda_0 \\ &\quad + g(x^*(\lambda_0))'(\lambda-\lambda_0) = f(x^*(\lambda_0)) + g(x^*(\lambda_0))'\lambda \end{aligned}$$

STRONG DUALITY IN CONVEX PROGRAMMING

- Consider the **convex programming** problem

$$\begin{array}{rl} \min_x & f(x) \\ \text{s.t.} & g_i(x) \leq 0,\ i \in I \\ & A_j x = b_j,\ j \in E \end{array} \qquad I \cup E = \{1, \ldots, m\}$$

where are f, g_i are convex functions.

- We say that **Slater's constraint qualification** is verified if the problem is strictly feasible:

$$\exists x:\ g_i(x) < 0,\ \forall i \in I,\ A_j x = b_j,\ \forall j \in E$$

- **Strong duality** always holds if Slater's constraint qualification is satisfied

- Other types of constraint qualifications exist

DUALITY AND KKT CONDITIONS FOR CONVEX PROBLEMS

THEOREM

Let x^* be the solution of a convex programming problem and f, g_i differentiable at x^*. Any λ^* satisfying the KKT conditions with x^* solves the dual problem.

Proof:

- Assume x^*, λ^* satisfy the KKT conditions and consider

$$\mathcal{L}(x, \lambda^*) = f(x) + \sum_{i \in I} \lambda_i^* g_i(x) + \sum_{j \in E} \lambda_j^* (A_j x - b_j)$$

- $\mathcal{L}(x, \lambda^*)$ is differentiable w.r.t. x at x^*, and is also a convex function of x, as $\lambda_i^* \geq 0$ for all $i \in I$

- By convexity of $\mathcal{L}(x, \lambda^*)$ we obtain

$$\mathcal{L}(x, \lambda^*) \geq \mathcal{L}(x^*, \lambda^*) + \overbrace{\nabla_x \mathcal{L}(x^*, \lambda^*)'}^{=0 \text{ because of KKT}} (x - x^*) = \mathcal{L}(x^*, \lambda^*)$$

DUALITY AND KKT CONDITIONS FOR CONVEX PROBLEMS

- Since $\mathcal{L}(x, \lambda^*) \geq \mathcal{L}(x^*, \lambda^*)$ for all x we get

$$\begin{aligned} q(\lambda^*) &= \inf_x \mathcal{L}(x, \lambda^*) = \mathcal{L}(x^*, \lambda^*) \\ &= f(x^*) + \sum_{i \in I} \underbrace{\lambda_i^* g_i(x^*)}_{=0 \text{ (complementarity)}} + \sum_{j \in E} \lambda_j^* (\underbrace{A_j x^* - b_j}_{=0 \text{ (feasibility)}}) = f(x^*) \end{aligned}$$

- Since $q(\lambda) \leq f(x^*)$ for all dual feasible λ, it follows that

$$q(\lambda) \leq q(\lambda^*)$$

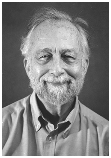
- As λ^* is dual feasible, it is therefore an optimizer of the dual problem. □

- Note that we have also proved that the duality gap is zero, as $q(\lambda^*) = f(x^*)$

WOLFE'S DUAL PROBLEM

- **Wolfe's dual problem** is defined as follows:

$$\begin{array}{rl} \max_{x,\lambda} & \mathcal{L}(x,\lambda) \\ \text{s.t.} & \nabla_x \mathcal{L}(x,\lambda) = 0 \\ & \lambda_i \geq 0,\ \forall i \in I \end{array}$$



Philip S. Wolfe
(1927–2016)

THEOREM

Consider a convex programming problem with f, g_i differentiable on $\mathbb{R}^n$.
Let x^*, λ^* satisfy the KKT conditions and LICQ hold.
Then x^*, λ^* is an optimizer of Wolfe's dual problem.

Proof:

- Since (x^*, λ^*) satisfies the KKT conditions it is a feasible point of Wolfe's dual problem, and moreover $\mathcal{L}(x^*, \lambda^*) = f(x^*)$
- For any (x, λ) satisfying $\nabla_x \mathcal{L}(x, \lambda) = 0$, $\lambda_i \geq 0$, $\forall i \in I$, we get

$$
\begin{aligned}
\mathcal{L}(x^*, \lambda^*) &= f(x^*) \geq f(x^*) + \sum_{i \in I} \overbrace{\lambda_i g_i(x^*)}^{\leq 0} + \sum_{j \in E} \lambda_j \overbrace{(A_j x^* - b_j)}^{= 0} \\
&= \underbrace{\mathcal{L}(x^*, \lambda) \geq \mathcal{L}(x, \lambda) + \overbrace{\nabla_x \mathcal{L}(x, \lambda)'}^{= 0}(x^* - x)}_{\text{convexity of } \mathcal{L}(x,\lambda)} \\
&= \mathcal{L}(x, \lambda)
\end{aligned}
$$

- Hence $\mathcal{L}(x^*, \lambda^*) = f(x^*)$ is the maximum achievable value of $\mathcal{L}(x, \lambda)$ under the constraints $\nabla_x \mathcal{L}(x, \lambda) = 0$, $\lambda_i \geq 0$, $\forall i \in I$. $\square$

DUAL LINEAR PROGRAM

- Consider the linear program

$$\begin{array}{rl} \min_x & c'x \\ \text{s.t.} & Ax \leq b \end{array}$$

- The dual function is

$$q(\lambda) = \inf_x \{c'x + \lambda'(Ax - b)\} = \inf_x \{(c + A'\lambda)'x - b'\lambda\}$$

- $q(\lambda) > -\infty$ only when $c + A'\lambda = 0$, and $q(\lambda) = -b'\lambda$

- The dual problem is therefore

$$\begin{array}{rl} \max_\lambda & -b'\lambda \\ \text{s.t.} & A'\lambda = -c \\ & \lambda \geq 0 \end{array} \quad \Rightarrow \quad \begin{array}{rl} \min_\lambda & b'\lambda \\ \text{s.t.} & A'\lambda = -c \\ & \lambda \geq 0 \end{array}$$

- It is easy to prove that the dual of the dual LP is the original LP ($\min_{x,s} c'x$ s.t. $Ax + s = b$, $s \geq 0$). The original x = dual vector of constraint $-A'\lambda + c = 0$, and s = dual vector of constraint $\lambda \geq 0$.

THEOREM OF ALTERNATIVES

THEOREM (THEOREM OF ALTERNATIVES)

For given $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, exactly one of the following two alternatives is true:

1. *there exists x such that $Ax \leq b$*
2. *there exists y such that $y \geq 0$, $A'y = 0$, $b'y < 0$*

LEMMA (FARKAS' LEMMA)

For a given matrix A and vector b, exactly one of the following two alternatives is true:

1. *there exists x such that $Ax = b$, $x \geq 0$*
2. *there exists y such that $A'y \geq 0$, $b'y < 0$*

Gyula Farkas
(1847–1930)

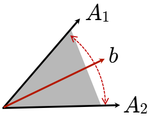
GEOMETRIC INTERPRETATION

SUPPLEMENTARY MATERIAL

Farkas' lemma has the following geometric interpretation.
Let A_i be the ith column of A, $i = 1, \ldots, n$, $A = [A_1 \ A_2 \ \ldots \ A_n]$

- **1st alternative**:

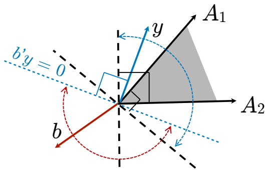
$$b = \sum_{i=1}^{n} x_i A_i,\ x_i \geq 0,\ i = 1, \ldots, n$$



b is in the convex cone generated by the columns of A

- **2nd alternative**:

$$\begin{aligned} y' A_i &\geq 0,\ i = 1, \ldots, n \\ y' b &< 0 \end{aligned}$$



vector b cannot be in the convex cone generated by the columns of A

DUAL LINEAR PROGRAM

THEOREM (STRONG LP DUALITY)

1. *If either the primal or the dual LP has a finite solution, so does the other and* $c'x^* = -b'\lambda^*$ (***strong duality***)
2. *If one of the two is **unbounded** the other is **infeasible***

- To see that infeasibility of dual LP implies unboundedness of a feasible primal LP, apply Farkas' Lemma with matrices $-A'$, c

$$-A'\lambda = c,\ \lambda \geq 0 \text{ infeasible} \quad \Longrightarrow \quad \exists d \in \mathbb{R}^n:\ -Ad \geq 0,\ c'd < 0$$

- Take a feasible $x_0 \in \mathbb{R}^n$. Then $A(x_0 + \sigma d) = Ax_0 + \sigma Ad \leq b, \forall \sigma \geq 0$, and $c'(x_0 + \sigma d) = c'x_0 - \sigma|c'd|$

- As σ can be arbitrarily large, the infimum of the primal LP is $-\infty$.

DUAL LP WITH NONNEGATIVE VARIABLES

SUPPLEMENTARY MATERIAL

- Consider the linear program

$$\begin{array}{rl} \min_x & c'x \\ \text{s.t.} & Ax \geq b \\ & x \geq 0 \end{array}$$

- The dual function is

$$q(\lambda,\nu) = \inf_x \{c'x + \lambda'(b - Ax) - \nu' x\} = \inf_x \{(c - A'\lambda - \nu)'x + b'\lambda\} = b'\lambda$$

for $c - A'\lambda - \nu = 0, \nu \geq 0$, or equivalently $A'\lambda \leq c$

- The dual problem is therefore

$$\begin{array}{rl} \max_\lambda & b'\lambda \\ \text{s.t.} & A'\lambda \leq c \\ & \lambda \geq 0 \end{array}$$

- At optimality $c'x^* = b'\lambda^*$

DUAL LP AND LINEAR COMPLEMENTARITY PROBLEM (LCP)

- A **linear complementarity problem** (LCP) is a **feasibility problem** of the form

(Cottle, Pang, Stone, 2009)

$$\begin{aligned} w &= Mz + q \\ w'z &= 0 \\ w, z &\geq 0 \end{aligned}$$

- By introducing the vector s of slack variables, $s = Ax - b \geq 0$, the KKT conditions for the following LP are

$$\begin{array}{rl} \min_x & c'x \\ \text{s.t.} & Ax \geq b \\ & x \geq 0 \end{array} \quad \Longrightarrow \quad \begin{aligned} & c - A'\lambda - \nu = 0 \\ & Ax - b - s = 0 \\ & x, \lambda, \nu, s \geq 0 \\ & x'\nu = \lambda' s = 0 \end{aligned}$$

- Therefore, the original LP can be solved by solving the LCP

$$\begin{bmatrix} \nu \\ s \end{bmatrix} = \begin{bmatrix} 0 & -A' \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} + \begin{bmatrix} c \\ -b \end{bmatrix}, \ \left[\begin{smallmatrix} x \\ \lambda \end{smallmatrix}\right], \left[\begin{smallmatrix} \nu \\ s \end{smallmatrix}\right] \geq 0, \ x'\nu = \lambda' s = 0$$

DUAL QUADRATIC PROGRAM

- Consider the quadratic program

$$\begin{array}{ll} \min_x & \frac{1}{2}x'Qx + c'x \\ \text{s.t.} & Ax \leq b \end{array} \qquad Q = Q' \succ 0$$

- The dual function is $q(\lambda) = \inf_x \left\{ \frac{1}{2}x'Qx + c'x + \lambda'(Ax - b) \right\}$

- Since $Q \succ 0$ the infimum is achieved when $0 = \nabla_x \mathcal{L}(x, \lambda) = Qx + c + A'\lambda$, i.e., for $x = -Q^{-1}(c + A'\lambda)$.

- By substitution, Lagrange's dual QP problem is therefore

$$\max_{\lambda \geq 0} -\left(\frac{1}{2}\lambda'(AQ^{-1}A')\lambda + (b + AQ^{-1}c)'\lambda + \frac{1}{2}c'Q^{-1}c \right)$$

DUAL QP AND LCP

- Let $Q \succ 0$ and consider the dual QP problem

$$\begin{array}{ll} \min & \frac{1}{2}\lambda'(AQ^{-1}A')\lambda + (b + AQ^{-1}c)'\lambda \\ \text{s.t.} & \lambda \geq 0 \end{array}$$

- The KKT conditions for the dual QP are the **LCP problem**

$$\begin{aligned} H\lambda + d &= s \\ s'\lambda &= 0 \\ s, \lambda &\geq 0 \end{aligned}$$

where $H = AQ^{-1}A'$ is the dual Hessian and $d = b + AQ^{-1}c$

- We can therefore solve the QP problem as an LCP to get the dual solution λ^* and then reconstruct the primal solution $x^* = -Q^{-1}(c + A'\lambda^*)$

LCP AND DUAL QP

- Vice versa, let $M = M' \succ 0$, $M \in \mathbb{R}^{n\times n}$, and consider the LCP

$$\begin{array}{c} x = My + d \\ 0 \le x \perp y \ge 0 \end{array}$$

- Consider the QP problem

$$\begin{array}{rl} \min & \frac{1}{2}y'My + d'y \\ \text{s.t.} & y \ge 0 \end{array}$$

- The corresponding KKT optimality conditions are

$$\begin{array}{rcll} My + d - x & = & 0 & \\ y & \ge & 0 & \\ x & \ge & 0 & \\ x_i y_i & = & 0, & i = 1, \ldots, n \end{array}$$

that are exactly the given LCP

WOLFE'S DUAL QP

SUPPLEMENTARY MATERIAL

- Consider now Wolfe's dual problem

$$\begin{array}{rl} \max_{x,\lambda} & \frac{1}{2}x'Qx + c'x + \lambda'(Ax-b) \\ \text{s.t.} & Qx + c + A'\lambda = 0,\ \lambda \geq 0 \end{array}$$

- We can subtract $0 = (Qx + c + A'\lambda)'x$ without changing the function and get the convex programming problem

$$\begin{array}{rl} \max_{x,\lambda} & -\frac{1}{2}x'Qx - \lambda'b \\ \text{s.t.} & Qx + c + A'\lambda = 0 \\ & \lambda \geq 0 \end{array}$$

- Note that Wolfe's dual QP only requires $Q \succeq 0$.

DUAL OF QP REFORMULATION OF LASSO

- Consider again the LASSO problem

$$\min_x \frac{1}{2}\|Ax-b\|_2^2+\gamma\|x\|_1 \qquad A\in\mathbb{R}^{m\times n},\ b\in\mathbb{R}^m,\ \gamma>0$$

- With $x=y-z$ and $y,z\geq 0$, LASSO becomes the positive semidefinite QP

$$\min_{y,z\geq 0} \frac{1}{2}\|A(y-z)-b\|_2^2+\gamma 1'(y+z)$$

where $1'=[1\ \ldots\ 1]$ (as $\gamma>0$ at least one of y_i^*, z_i^* will be zero at optimality)

- The above QP is the dual of the following **least distance programming** (LDP) (constrained LS) problem (see next slide)

$$\begin{array}{ll} \min_v & \frac{1}{2}\|v-b\|_2^2 - b'b \\ \text{s.t.} & \|A'v\|_\infty \leq \gamma \end{array}$$

DUAL OF QP REFORMULATION OF LASSO

- Proof: The constrained LS problem is equivalent to the following QP

$$\begin{array}{rl} \min_v & \frac{1}{2}v'v - b'v - \frac{1}{2}b'b \\ \text{s.t.} & -\gamma 1 \leq A'v \leq \gamma 1 \end{array}$$

whose dual QP problem is exactly the original LASSO's QP reformulation

$$\min_{y,z\geq 0} \frac{1}{2}\left[\begin{smallmatrix} y \\ z \end{smallmatrix}\right]\left[\begin{smallmatrix} A' \\ -A' \end{smallmatrix}\right] I^{-1}\left[\begin{smallmatrix} A & -A \end{smallmatrix}\right]\left[\begin{smallmatrix} y \\ z \end{smallmatrix}\right] + (\gamma\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] - \left[\begin{smallmatrix} A' \\ -A' \end{smallmatrix}\right] I^{-1}b)'\left[\begin{smallmatrix} y \\ z \end{smallmatrix}\right] + \frac{1}{2}b'b - \frac{1}{2}b'b$$

$\square$

- The LDP reformulation of LASSO is always a **strictly convex** QP with m variables, $2n$ constraints, and Hessian = identity matrix

- The original QP formulation is only **convex** with $2n$ variables and $2n$ constraints

SUPPORT VECTOR REGRESSION

(Smola, Schölkopf, 2004)

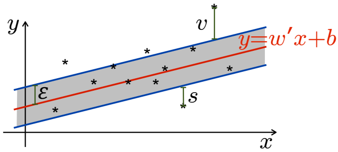
- We have a training set $(x_1, y_1), \ldots, (x_N, y_N), x_i \in \mathbb{R}^n, y \in \mathbb{R}$ and want to fit a linear function

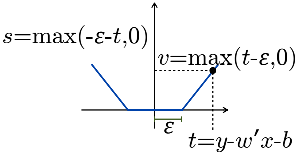
$$f(x) = w'x + b \quad w \in \mathbb{R}^n, b \in \mathbb{R}$$

such that each $|y_i - f(x_i)| \leq \epsilon$

- Since such a function f may not exist, we want to penalize $|y_i - f(x_i)| > \epsilon$

$$\begin{array}{ll} \min\limits_{w,b,v,s} & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^N (v_i + s_i) \\ \text{s.t.} & y_i - w'x_i - b \leq \epsilon + v_i \\ & y_i - w'x_i - b \geq -\epsilon - s_i \\ & v_i, s_i \geq 0 \end{array}$$





SUPPORT VECTOR REGRESSION

SUPPLEMENTARY MATERIAL

- By setting $X = [x_1 \ \dots \ x_N], Y = [y_1 \ \dots \ y_N]'$, we can rewrite in vector form

$$\begin{aligned} \min_{w,b,v,s} \quad & \frac{1}{2}w'w + C\,\mathbb{1}'(v+s) \\ \text{s.t.} \quad & Y - X'w - b\,\mathbb{1} \leq \epsilon\,\mathbb{1} + v \\ & Y - X'w - b\,\mathbb{1} \geq -\epsilon\,\mathbb{1} - s \\ & v, s \geq 0 \end{aligned}$$

- Introduce the vectors of $\mathbb{R}^N$ of Lagrange multipliers $\alpha, \beta, \gamma, \delta \geq 0$

- The Lagrangian function is

$$\begin{aligned} \mathcal{L}\left(\begin{bmatrix} w \\ b \\ v \\ s \end{bmatrix}, \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix}\right) = & \frac{1}{2}w'w + C\,\mathbb{1}'(v+s) + \alpha'(Y - X'w - (b+\epsilon)\,\mathbb{1} - v) \\ & + \beta'(-Y + X'w + (b-\epsilon)\,\mathbb{1} - s) - \gamma'v - \delta's \end{aligned}$$

- The dual function $q(\alpha, \beta, \gamma, \delta) = \inf_{w,b,v,s} \mathcal{L}(w, b, v, s, \alpha, \beta, \gamma, \delta)$

- Let us zero the partial derivatives of $\mathcal{L}$ with respect to w, b, v, s:

$$\begin{array}{rcll}
0 = \frac{\partial \mathcal{L}}{\partial w} & = & w - X\alpha + X\beta & \Rightarrow \quad w = X(\alpha - \beta) \\
0 = \frac{\partial \mathcal{L}}{\partial b} & = & -\alpha' \mathbb{1} + \beta' \mathbb{1} & \Rightarrow \quad \mathbb{1}'(\alpha - \beta) = 0 \\
0 = \frac{\partial \mathcal{L}}{\partial v} & = & C \mathbb{1} - \alpha - \gamma & \Rightarrow \quad \gamma = C \mathbb{1} - \alpha \geq 0 \\
0 = \frac{\partial \mathcal{L}}{\partial s} & = & C \mathbb{1} - \beta - \delta & \Rightarrow \quad \delta = C \mathbb{1} - \beta \geq 0
\end{array}$$

- By substituting the above expressions in the Lagrangian we get

$$\begin{array}{rcl}
q(\alpha, \beta, \gamma, \delta) & = & \frac{1}{2} w'w + (Y - X'w)'(\alpha - \beta) - \epsilon \mathbb{1}'(\alpha + \beta) \\
& = & -\frac{1}{2}(\alpha - \beta)'X'X(\alpha - \beta) + Y'(\alpha - \beta) - \epsilon \mathbb{1}'(\alpha + \beta)
\end{array}$$

- The dual problem is therefore the following QP

$$\begin{array}{rl}
\min_{\alpha,\beta} & \frac{1}{2}(\alpha - \beta)'X'X(\alpha - \beta) - Y'(\alpha - \beta) + \epsilon \mathbb{1}'(\alpha + \beta) \\
\text{s.t.} & 0 \leq \alpha \leq C \mathbb{1}, \quad 0 \leq \beta \leq C \mathbb{1}, \quad \mathbb{1}'(\alpha - \beta) = 0
\end{array}$$

SUPPLEMENTARY MATERIAL

- After solving the dual QP problem we can retrieve

$$\begin{aligned} w &= X(\alpha^* - \beta^*) = \sum_{i=1}^N (\alpha_i^* - \beta_i^*) x_i \\ f(x) &= w'x + b = (\alpha^* - \beta^*)' X' x + b = \sum_{i=1}^N (\alpha_i^* - \beta_i^*) x_i' x + b \end{aligned}$$

$$\boxed{f(x) = \sum_{i=1}^N (\alpha_i^* - \beta_i^*) x_i' x + b}$$

(see next slide for how to reconstruct b)

- $f(x)$ is defined by a linear combination of the training vectors x_i
- The vectors x_i for which $\alpha_i^* - \beta_i^* \neq 0$ are called **support vectors**
- Note that the QP is also equivalent to the ℓ_1-regularized problem

$$\begin{array}{ll} \min_z & \frac{1}{2} z' X' X z - Y' z + \epsilon \|z\|_1 \\ \text{s.t.} & |z_i| \leq C, \quad \sum_{i=1}^N z_i = 0 \end{array}$$

SUPPLEMENTARY MATERIAL

- The scalar b can be retrieved from the complementarity slackness conditions

$$
\begin{array}{rcl}
0 & = & \alpha_i(y_i - x_i'w - (b+\epsilon) - v_i), \quad i = 1, \ldots, N \\
0 & = & \beta_i(-y_i + x_i'w + (b-\epsilon) - s_i) \\
0 & = & \gamma_i v_i = (C - \alpha_i)v_i \\
0 & = & \delta_i s_i = (C - \beta_i)s_i
\end{array}
$$

- if any $\alpha_i^* \in (0, C)$ then $v_i^* = 0 \Rightarrow b^* = y_i - x_i'w^* - \epsilon$

- if any $\beta_i^* \in (0, C)$ then $s_i^* = 0 \Rightarrow b^* = y_i - x_i'w^* + \epsilon$

SUPPLEMENTARY MATERIAL

- Otherwise, consider the case all $\alpha_i^*, \beta_i^* \in \{0, C\}$
- α_i^*, β_i^* cannot be positive at the same time, as they refer to bilateral constraints ($y_i - w'x_i - b$ cannot be both positive and negative)

$$\begin{array}{rclcl}
\alpha_i = 0 & \Rightarrow & v_i = 0 & \Rightarrow & y_i - x_i'w - (b+\epsilon) \leq 0 \\
\beta_i = 0 & \Rightarrow & s_i = 0 & \Rightarrow & -y_i + x_i'w + (b-\epsilon) \leq 0 \\
\alpha_i = C & \Rightarrow & \beta_i = 0 & \Rightarrow & s_i = 0, \quad -y_i + x_i'w + (b-\epsilon) \leq 0 \\
\beta_i = C & \Rightarrow & \alpha_i = 0 & \Rightarrow & v_i = 0, \quad y_i - x_i'w - (b+\epsilon) \leq 0
\end{array}$$

- Let $\mathcal{I} = \{i : \alpha_i^* = 0 \quad \text{or} \quad \beta_i^* = C\}$ and $\mathcal{J} = \{i : \alpha_i^* = C \quad \text{or} \quad \beta_i^* = 0\}$. Then

$$\begin{aligned}
b^* &\geq y_i - x_i'w^* - \epsilon, \quad \forall i \in \mathcal{I} \\
b^* &\leq y_i - x_i'w^* + \epsilon, \quad \forall i \in \mathcal{J}
\end{aligned}$$

- Therefore, any $b^* \in [\max_{i\in\mathcal{I}}\{y_i - x_i'w^* - \epsilon\}, \min_{i\in\mathcal{J}}\{y_i - x_i'w^* + \epsilon\}]$ is optimal

SUPPLEMENTARY MATERIAL

- **Kernel trick**: if we generalize x_i to an arbitrary nonlinear basis $\phi(x_i)$ we get

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \beta_i^*) k(x_i, x) + b$$

where $k(x,y) = \phi'(x)\phi(y)$ is a **kernel function**, $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

- Example: $x \in \mathbb{R}^2$, $\phi(x) = [x_1^2 \quad \sqrt{2}x_1x_2 \quad x_2^2]'$, $k(x,y) = (x'y)^2$
- The (i,j)th term $x_i'x_j$ of the dual Hessian gets replaced by $k(x_i, x_j)$
- b depends on $x_i'w = x_i'X(\alpha - \beta)$ that gets replaced by $k(x_i, X)(\alpha^* - \beta^*)$
- Therefore ϕ, w are not required, and can have arbitrary dimensions !
- Example: **Gaussian radial basis function kernel** $k(x,y) = e^{-\frac{1}{2}\|x-y\|^2/\sigma^2}$ (RBF) the corresponding ϕ is infinite dimensional

EXAMPLE OF SUPPORT VECTOR REGRESSION

- Generate $N = 100$ random samples of the course-logo function

$$f(x_1, x_2) = -e^{-(x_1^2+x_2^2)} + 0.3\sin\left(\frac{1}{10}x_1^3 + x_2^2\right) + 1.2$$

- Solve SVR problem with $C = 100$, $\epsilon = 0.01$, Gaussian kernel with $\sigma = 1$