Optimization and Data Science

Lecture 23: Active Set Method

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Contents

Recall: Lagrange multiplier rule for inequality constraints: KKT system

- The first order necessary optimality condition for a local solution x^* to

$$\min_{x\in\mathbb{R}^n} f(x) \;\text{ subject to } \begin{cases} g(x) \le 0 \\ h(x) = 0 \end{cases}$$

is that $(x^*, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m_{\geq 0}$ solves the Karush-Kuhn-Tucker (KKT) system

$$\begin{aligned} \nabla_x L(x^*\lambda, \mu) = \nabla f(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{j=1}^{m} \mu_j \nabla g_j(x^*) &= 0 \\ \mu^\top g(x^*) &= 0, \\ h(x^*) &= 0, \\ g(x^*) &\le 0, \end{aligned}$$

with the Lagrangian

$$L(x, \lambda, \mu) := f(x) + \lambda^\top h(x) + \mu^\top g(x) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j g_j(x).$$

Contents

Recall: Lagrange multiplier rule and KKT system

- In the problem with equality constraints only,

$$\min_{x\in\mathbb{R}^n} f(x) \;\text{ subject to } h(x)=0,$$

the solution $(x^*,\lambda)\in\mathbb{R}^n\times\mathbb{R}^p$ of the system

$$\nabla_{(x,\lambda)}L(x^*,\lambda)=0 \quad\Leftrightarrow\quad \left\{\begin{array}{rcrcl} \nabla_x L(x^*,\lambda) & = & \nabla f(x^*)+\sum_{i=1}^{p}\lambda_i\nabla h_i(x^*) & = & 0\\ \nabla_\lambda L(x^*,\lambda) & = & h(x^*) & = & 0 \end{array}\right.$$

can be computed by Newton or Quasi-Newton methods, leading to the SQP method.

Newton-SQP step for equality constraints

- Newton-SQP step:

$$\nabla^2_{(x,\lambda)} L(x_k,\lambda_k) \begin{pmatrix} d \\ \delta \end{pmatrix} = -\nabla_{(x,\lambda)} L(x_k,\lambda_k).$$

or

$$\begin{pmatrix} \nabla^2_{xx} L(x_k,\lambda_k) & \nabla^2_{\lambda x} L(x_k,\lambda_k) \\ \nabla^2_{x\lambda} L(x_k,\lambda_k) & \nabla^2_{\lambda\lambda} L(x_k,\lambda_k) \end{pmatrix} \begin{pmatrix} d \\ \delta \end{pmatrix} = -\begin{pmatrix} \nabla_x L(x_k,\lambda_k) \\ \nabla_\lambda L(x_k,\lambda_k) \end{pmatrix}$$

or

$$\begin{pmatrix} \nabla^2_{xx} L(x_k,\lambda_k) & \nabla h_1(x_k) \cdots \nabla h_p(x_k) \\ \nabla h_1(x_k)^\top & \\ \vdots & 0 \\ \nabla h_p(x_k)^\top & \end{pmatrix} \begin{pmatrix} d \\ \delta \end{pmatrix} = -\begin{pmatrix} \nabla_x L(x_k,\lambda_k) \\ h(x_k) \end{pmatrix}$$

A different interpretation of the SQP method for equality constraints

- Approximate L in the vicinity of the current iterate (x_k, λ_k) by the quadratic function

$$L(x_k + d, \lambda_k + \delta) \approx L(x_k, \lambda_k) + \begin{pmatrix} \nabla_x L(x_k, \lambda_k) \\ \nabla_\lambda L(x_k, \lambda_k) \end{pmatrix}^\top \begin{pmatrix} d \\ \delta \end{pmatrix} + \frac{1}{2} \begin{pmatrix} d \\ \delta \end{pmatrix}^\top \nabla^2_{(x,\lambda)} L(x_k, \lambda_k) \begin{pmatrix} d \\ \delta \end{pmatrix}$$

$$= \frac{1}{2} \begin{pmatrix} d \\ \delta \end{pmatrix}^\top A \begin{pmatrix} d \\ \delta \end{pmatrix} + b^\top \begin{pmatrix} d \\ \delta \end{pmatrix} + c.$$

- Minimize this approximation w.r.t. (d, δ). Necessary optimality condition:

$$A \begin{pmatrix} d \\ \delta \end{pmatrix} + b = \nabla^2_{(x,\lambda)} L(x_k, \lambda_k) \begin{pmatrix} d \\ \delta \end{pmatrix} + \nabla_{(x,\lambda)} L(x_k, \lambda_k) = 0$$

- This gives (d, δ) as solution of

$$\nabla^2_{(x,\lambda)} L(x_k, \lambda_k) \begin{pmatrix} d \\ \delta \end{pmatrix} = -\nabla_{(x,\lambda)} L(x_k, \lambda_k).$$

$\rightsquigarrow$ (d, δ) is the Newton direction.

Quadratic approximation of the Lagrangian

$$
\begin{aligned}
L_k(d,\delta) &= \frac{1}{2}\begin{pmatrix} d \\ \delta \end{pmatrix}^\top \begin{pmatrix} \nabla^2_{xx}L & \nabla^2_{\lambda x}L \\ \nabla^2_{x\lambda}L & 0 \end{pmatrix}\begin{pmatrix} d \\ \delta \end{pmatrix} + \begin{pmatrix} \nabla_x L \\ \nabla_\lambda L \end{pmatrix}^\top \begin{pmatrix} d \\ \delta \end{pmatrix} + L(x_k,\lambda_k) \\
&= \frac{1}{2}\begin{pmatrix} d \\ \delta \end{pmatrix}^\top \begin{pmatrix} \nabla^2_{xx}Ld + \nabla^2_{\lambda x}L\delta \\ \nabla^2_{x\lambda}Ld \end{pmatrix} + \nabla_x L^\top d + h(x_k)^\top \delta + L(x_k,\lambda_k) \\
&= \frac{1}{2}\left(d^\top \nabla^2_{xx}Ld + d^\top \nabla^2_{\lambda x}L\delta + \delta^\top \nabla^2_{x\lambda}Ld\right) + \nabla_x L^\top d + h(x_k)^\top \delta + L(x_k,\lambda_k) \\
&= \frac{1}{2}d^\top \nabla^2_{xx}Ld + \nabla_x L^\top d + L(x_k,\lambda_k) + \delta^\top \nabla^2_{x\lambda}Ld + \delta^\top h(x_k).
\end{aligned}
$$

$\rightsquigarrow$ L_k is the Lagrangian of the following problem with multiplier δ:

$$
\min_{d\in\mathbb{R}^n} \frac{1}{2}d^\top \nabla^2_{xx}Ld + \nabla_x L^\top d + L(x_k,\lambda_k) \quad \text{s.t.} \quad \nabla^2_{x\lambda}Ld + h(x_k) = 0.
$$

Quadratic approximation of the Lagrangian

- L_k is the Lagrangian of the problem

$$\min_{d\in\mathbb{R}^n} \frac{1}{2} d^\top \nabla^2_{xx} L d + \nabla_x L^\top d + L(x_k, \lambda_k) \quad \text{s.t. } \nabla^2_{x\lambda} L d + h(x_k) = 0.$$

- Using

$$\nabla^2_{x\lambda} L(x,\lambda) = \left(\nabla h_i(x)^\top\right)_{i=1}^{p} \in \mathbb{R}^{p\times n},$$
$$\nabla^2_{x\lambda} L(x,\lambda) d = \left(\nabla h_i(x)^\top d\right)_{i=1}^{p} \in \mathbb{R}^{p},$$

- ... we re-write the constraint as

$$\nabla h_i(x_k)^\top d + h_i(x_k) = 0, \quad i = 1, \dots, p.$$

- This is a linearization of the original constraint functions h_i around x_k:

$$h_i(x_k + d) = h_i(x_k) + \nabla h_i(x_k)^\top d + \mathcal{O}(\|d\|^2).$$

Different interpretation of SQP: Summary

Algorithm (SQP method – rewritten):

1. Choose initial guess $(x_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^p$.
2. For $k = 0, 1, \ldots$:
    1. Approximate the original problem by a quadratic problem with linear constraints, using a Taylor approximation around (x_k, λ_k).
    2. Solve this approximative problem $\rightsquigarrow (d_k, \delta_k)$.
    3. Choose an efficient step-size $\rho_k > 0$.
    4. Update

$$(x_{k+1}, \lambda_{k+1}) = (x_k, \lambda_k) + \rho_k(d_k, \delta_k),$$

until a stopping criterion is satisfied.

Contents

Extension to inequality constraints

- We apply the same idea for a problem with additional inequality constraints:

$$\min_{x\in\mathbb{R}^n} f(x) \text{ subject to } \begin{cases} g(x) \le 0 \\ h(x) = 0. \end{cases}$$

- The computations are the same.
- First order necessary condition: $(x^*, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m_{\ge 0}$ solves the KKT system

$$\begin{aligned} \nabla_x L(x^*\lambda, \mu) = \nabla f(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{j=1}^{m} \mu_j \nabla g_j(x^*) &= 0 \\ \mu^\top g(x^*) &= 0, \\ h(x^*) &= 0, \\ g(x^*) &\le 0, \end{aligned}$$

with the Lagrangian

$$L(x, \lambda, \mu) := f(x) + \lambda^\top h(x) + \mu^\top g(x) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j g_j(x).$$

Quadratic approximation of the Lagrangian with inequality constraints

- We have additional multipliers $\mu \in \mathbb{R}^m_{\geq 0}$ now.
- Approximate L by the quadratic function

$$
\begin{aligned}
L(x_k + d, \lambda_k + \delta, \mu_k + \gamma) &\approx L + \nabla_{(x,\lambda,\mu)} L^\top \begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix} + \frac{1}{2} \begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix}^\top \nabla^2_{(x,\lambda,\mu)} L \begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix} \\
&= \frac{1}{2} \begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix}^\top A \begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix} + b^\top \begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix} + c =: L_k(d, \delta, \gamma),
\end{aligned}
$$

  where on the right-hand side L always means $L(x_k, \lambda_k, \mu_k)$.
- Minimize this approximation w.r.t. (d, δ, γ).
- Have to take into account the inequality constraints and complementarity condition.

$\rightsquigarrow$ Direct solution via the necessary condition is not possible.

Quadratic approximation of the Lagrangian

$$L_k(d,\delta) = \frac{1}{2}\begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix}^\top \begin{pmatrix} \nabla^2_{xx}L & \nabla^2_{\lambda x}L & \nabla^2_{\mu x}L \\ \nabla^2_{x\lambda}L & 0 & 0 \\ \nabla^2_{x\mu}L & 0 & 0 \end{pmatrix}\begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix} + \begin{pmatrix} \nabla_x L \\ \nabla_\lambda L \\ \nabla_\mu L \end{pmatrix}^\top \begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix} + L(x_k,\lambda_k,\mu_k)$$

$$= \frac{1}{2}\begin{pmatrix} d \\ \delta \\ \gamma \end{pmatrix}^\top \begin{pmatrix} \nabla^2_{xx}Ld + \nabla^2_{\lambda x}L\delta + \nabla^2_{\mu x}L\gamma \\ \nabla^2_{x\lambda}Ld \\ \nabla^2_{x\mu}Ld \end{pmatrix} + \nabla_x L^\top d + h(x_k)^\top\delta + g(x_k)^\top\gamma + L$$

$$= \frac{1}{2}\left(d^\top\nabla^2_{xx}Ld + d^\top\nabla^2_{\lambda x}L\delta + d^\top\nabla^2_{\mu x}L\gamma + \delta^\top\nabla^2_{x\lambda}Ld + \gamma^\top\nabla^2_{x\mu}L\gamma\right) + \nabla_x L^\top d + \ldots + L$$

$$= \frac{1}{2}d^\top\nabla^2_{xx}Ld + \nabla_x L^\top d + L + \delta^\top\nabla^2_{x\lambda}Ld + \gamma^\top\nabla^2_{x\mu}L\gamma + \delta^\top h(x_k) + \gamma^\top g(x_k).$$

$\rightsquigarrow$ This is the Lagrangian of the following problem with multipliers δ and γ:

$$\min_{d\in\mathbb{R}^n} \frac{1}{2}d^\top\nabla^2_{xx}Ld + \nabla_x L^\top d + L \quad \text{s.t.} \begin{cases} \nabla^2_{x\lambda}Ld + h(x_k) = 0 \\ \nabla^2_{x\mu}Ld + g(x_k) \le 0. \end{cases}$$

Quadratic approximation of the Lagrangian

- L_k is the Lagrangian of the problem

$$\min_{d\in\mathbb{R}^n} \frac{1}{2} d^\top \nabla^2_{xx} L d + \nabla_x L^\top d + L \quad \text{s.t.} \begin{cases} h_i(x_k) + \nabla h_i(x_k)^\top d = 0, i = 1, \dots, p \\ g_i(x_k) + \nabla g_i(x_k)^\top d \le 0, i = 1, \dots, m. \end{cases}$$

- Here we used again

$$\nabla^2_{x\lambda} L(x, \lambda, \mu) d = \left(\nabla h_i(x)^\top d\right)_{i=1}^{p}, \quad \nabla^2_{x\mu} L(x, \lambda, \mu) d = \left(\nabla g_i(x)^\top d\right)_{i=1}^{m}$$

to rewrite the constraints.

- These are again linearizations of the original constraint functions h and g around x_k.

$\rightsquigarrow$ Need a method to solve a quadratic problem with (affine-) linear equality and inequality constraints.

Contents

Quadratic cost function with linear equality + inequality constraints

- Solve a quadratic problem with (affine-) linear equality and inequality constraints:

$$\min_{d\in\mathbb{R}^n} \frac{1}{2} d^\top \nabla^2_{xx} L d + \nabla_x L^\top d + L \quad \text{s.t.} \quad \begin{cases} h_i(x_k) + \nabla h_i(x_k)^\top d = 0, i = 1, \dots, p, \\ g_i(x_k) + \nabla g_i(x_k)^\top d \leq 0, i = 1, \dots, m. \end{cases}$$

- Simplified notation:

$$\min_{d\in\mathbb{R}^n} \frac{1}{2} d^\top A d + b^\top d + c \ \text{ s.t. } \begin{cases} Hd + h & = & 0, \\ Gd + g & \leq & 0. \end{cases}$$

where

$$H = \left(\nabla h_i(x_k)^\top\right)_{i=1}^{p} = \left(\frac{\partial h_i}{\partial x_\ell}(x_k)\right)_{\substack{i=1,\dots,p \\ \ell=1,\dots,n}}, G = \left(\nabla g_i(x_k)^\top\right)_{i=1}^{m} = \left(\frac{\partial g_i}{\partial x_\ell}(x_k)\right)_{\substack{i=1,\dots,m \\ \ell=1,\dots,n}}$$

$$h = h(x_k), g = g(x_k).$$

Active set strategy

- We want to solve

$$\min_{d\in\mathbb{R}^n} \frac{1}{2}d^\top A d + b^\top d + c \ \text{ s.t. } \begin{cases} Hd + h &= 0 \\ Gd + g &\leq 0. \end{cases}$$

- Lagrangian:

$$L(d,\lambda,\mu) = \frac{1}{2}d^\top A d + b^\top d + c + \underbrace{\lambda^\top(Hd+h)}_{(H^\top\lambda)^\top d + \lambda^\top h} + \underbrace{\mu^\top(Gd+g)}_{(G^\top\mu)^\top d + \mu^\top g}.$$

- KKT system:

$$\begin{aligned} Ad + b + H^\top\lambda + G^\top\mu &= 0 \\ Hd + h &= 0 \\ Gd + g &\leq 0 \\ \mu^\top(Gd+g) &= 0. \end{aligned}$$

Considering the active constraints only

- At given admissible d, we determine the active inequality constraints, i.e., the sets

$$\mathcal{A} := \{j \in \{1,\dots,m\} : (Gd)_j + g_j = 0\}, \quad \mathcal{I} := \{j \in \{1,\dots,m\} : (Gd)_j + g_j < 0\}. \quad (1)$$

- From the complementarity condition, we deduce:

$$\mu_j = 0, j \in \mathcal{I}.$$

$\rightsquigarrow$ We ignore the inactive constraints and compute a solution $(\hat{d}, \lambda, \hat{\mu}) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^{|\mathcal{A}|}$ of

$$\left.\begin{array}{rcl} A\hat{d} + H^\top \lambda + \hat{G}^\top \hat{\mu} & = & -(Ad + b) \\ H\hat{d} & = & 0 \\ \hat{G}\hat{d} & = & 0, \end{array}\right\} \quad (2)$$

where $\hat{G}$ contains the rows of G corresponding to the active constraints only:

$$\hat{G} = (G_{jk})_{j \in \mathcal{A}, k=1,\dots,n} \in \mathbb{R}^{|\mathcal{A}| \times n}.$$

Considering the active constraints only

- d is admissible: $Hd + h = 0,\ Gd + g \leq 0$.
- $(\hat{d}, \lambda, \hat{\mu})$ solves (2):

$$\begin{aligned} A\hat{d} + H^\top \lambda + \hat{G}^\top \hat{\mu} &= -(Ad + b) \\ H\hat{d} &= 0 \\ \hat{G}\hat{d} &= 0. \end{aligned}$$

- Set $(\mu_i)_{i \in \mathcal{A}} := \hat{\mu}, (\mu_i)_{i \in \mathcal{I}} := 0$. Then $\hat{G}^\top \hat{\mu} = G^\top \mu$.

$\rightsquigarrow$ $(d + \hat{d}, \lambda, \mu)$ solves

$$\begin{aligned} A(d + \hat{d}) + b + H^\top \lambda + G^\top \mu &= 0 \\ H(d + \hat{d}) + h &= 0 \\ G(d + \hat{d}) + g &\leq 0 \\ \mu^\top (G(d + \hat{d}) + g) &= 0. \end{aligned}$$

- Especially, $(d + \hat{d})$ is admissible again.
- If $\mu \geq 0$, we have found a point that satisfies the KKT conditions.

Algorithm: Active set strategy

1. Find an admissible point $d \in \mathbb{R}^n$.
2. For d, determine the set $\mathcal{A}$ of active inequality constraints defined in (1).
3. Compute a solution $(\hat{d}, \lambda, \hat{\mu})$ of the linear system (2).

(a) If $\|\hat{d}\| \leq \epsilon$:

- If $\hat{\mu} \geq 0$: stop, (d, λ, μ) with $(\mu_i)_{i \in \mathcal{A}} := \hat{\mu}, (\mu_i)_{i \in \mathcal{I}} := 0$ solves the KKT system.
- If there is $i \in \mathcal{A}$ with $\hat{\mu}_i < 0$: set $\mathcal{A} = \mathcal{A} \backslash \{i\}$ and go back to step 3.
  (Sensitivity Theorem: $\hat{\mu}_i < 0$: $(\hat{G}\hat{d})_i < 0 \Rightarrow f \downarrow$, inactive constraint will reduce cost.)

(b) If $\|\hat{d}\| > \epsilon$:

- Choose step-size ρ and set $d := d + \rho \hat{d}$.
  ($H\hat{d} = 0$, $\hat{G}\hat{d} = 0 \rightsquigarrow$ this will not violate the equality and the active inequality constraints.)
  Inactive constraints ($i \in \mathcal{I}$) must not be violated, too:

$$(G(d + \rho\hat{d}) + g)_i = \underbrace{(Gd)_i + g_i}_{\leq 0} + \rho(G\hat{d})_i \leq 0, i \in \mathcal{I} \quad \Rightarrow \quad \rho = \min_{i \in \mathcal{I}} \left\{ \frac{(Gd)_i + g_i}{(G\hat{d})_i} \right\}.$$

- Go back to step 2.

Contents

Support vector machine: optimization problam

- Already a quadratic problem with linear inequality constraints:

$$\min_{(a,b)\in\mathbb{R}^n\times\mathbb{R}} \underbrace{\|a\|_2^2}_{=\frac{1}{2}d^\top Ad} \quad \text{s.t.} \quad \underbrace{g_j(a,b) = \Delta - \left(a^\top z_j - b\right) f(z_j) \leq 0, j = 1,\dots,m}_{Gd+g\leq 0}$$

where

$$d = \begin{pmatrix} a \\ b \end{pmatrix}, \quad A = \begin{pmatrix} 2I_{n\times n} & 0_{n\times 1} \\ 0_{1\times n} & 0 \end{pmatrix},$$

$$g = \Delta, \quad f(z) := (f(z_j))_{j=1}^m \in \mathbb{R}^{n\times 1},$$

$$G_{ji} = \frac{\partial g_j}{\partial a_i}(a,b) = -f(z_j)z_{ji}, i = 1,\dots,n, \quad G_{j,n+1} = \frac{\partial g_j}{\partial b}(a,b) = f(z_j), j = 1,\dots,m,$$

$$G = \left(-\mathrm{diag}(f(z))z, f(z)\right), \quad z := (z_{ji})_{j=1,\dots,m,i=1,\dots,n} \in \mathbb{R}^{m\times n}.$$

$\leadsto$ Apply directly the active set algorithm.

Active set method for support vector machine

- Linear system (2) to be solved for $(\hat{d}, \hat{\mu}) \in \mathbb{R}^{n+1} \times \mathbb{R}^m$:

$$
\begin{aligned}
A\hat{d} + \hat{G}^\top \hat{\mu} &= -Ad \\
\hat{G}\hat{d} &= 0.
\end{aligned}
$$

- Inserting gives:

$$
\begin{pmatrix} 2I_{n\times n} & 0_{n\times 1} \\ 0_{1\times n} & 0 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} + \begin{pmatrix} -z^\top \mathrm{diag}(f(z)) \\ f(z)^\top \end{pmatrix} \hat{\mu} = -\begin{pmatrix} 2I_{n\times n} & 0_{n\times 1} \\ 0_{1\times n} & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}
$$

$$
\begin{pmatrix} -\mathrm{diag}(f(z))z, f(z) \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} 0_m \\ 0 \end{pmatrix}
$$

- ... or:

$$
\begin{aligned}
2\hat{a} - z^\top \mathrm{diag}(f(z))\hat{\mu} &= -2a \\
f(z)^\top \hat{\mu} &= 0 \\
-\mathrm{diag}(f(z))z\hat{a} + f(z)\hat{b} &= 0_m.
\end{aligned}
$$

What is important

- Problems with inequality constraints can be treated by Lagrange methods also.
- Since the KKT system contains inequalities, a special algorithm has to be used.
- Active set strategies are one important class of methods to solve such kind of problems.
- At first, an active set strategy approximates the nonlinear problem by a quadratic cost function with linearized constraints.
- This quadratic problem is then iteratively solved.
- The active set strategy starts with an admissible point.
- Then, the solution to the KKT system is updated by the solution of the homogenous problem considering the active inequality constraints only, and updating the active set.