Machine Learning

Lecture 11: Constrained Optimization

Prof. Dr. Aleksandar Bojchevski

29.05.24

Outline

Non-negative least squares

In many important use cases, negative weights would not be sensible, e.g., when fitting physical properties like resistance or density.

Hence, we need an additional constraint:

$$\begin{aligned} \text{minimize}_{\boldsymbol{w}} \quad & \mathcal{L}_{\mathrm{LS}}(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 \\ \text{subject to} \quad & w_i \geq 0 \quad \text{for } i = 1, \ldots, d\,. \end{aligned}$$

How do we solve this?

Optimization with inequality constraints

Constrained optimization problem: Given $f_0 : \mathbb{R}^d \to \mathbb{R}$ and $f_i : \mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}} && f_0(\boldsymbol{\theta}) \\ &\text{subject to} && f_i(\boldsymbol{\theta}) \leq 0 \quad \text{for } i = 1, \ldots, M\,. \end{aligned}$$

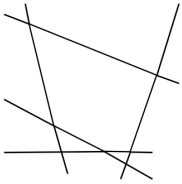
Feasibility: A point $\boldsymbol{\theta} \in \mathbb{R}^d$ is called feasible if and only if it satisfies the constraints of the optimization problem, i.e., $f_i(\boldsymbol{\theta}) \leq 0$ for all $i \in \{1, \ldots, M\}$.

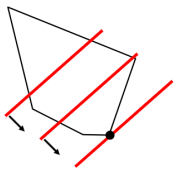
Minimum and minimizer: We call the optimal value the minimum p^*, and the point where the minimum is obtained the minimizer $\boldsymbol{\theta}^*$. Thus $p^* = f_0(\boldsymbol{\theta}^*)$.

Standard problems with inequality constraints

Linear Programming (LP)

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}} && \boldsymbol{c}^T\boldsymbol{\theta} \\ &\text{subject to} && \boldsymbol{A\theta} - \boldsymbol{b} \leq 0 \end{aligned}$$





The feasible set defines a convex polytope – a convex set resulting from the intersection of half spaces.

The optimum is always at a vertex of the polytope (assuming a unique solution).[1]

The simplex algorithm solves LPs by moving from vertex to vertex along edges.[2]

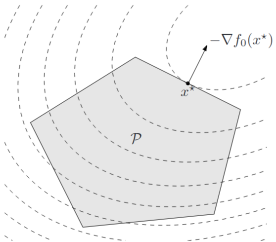
[1] If there are multiple solutions, $\boldsymbol{c}$ will be parallel to a face of the polytope.

[2] It can take exponential time in the dimension, but it is very efficient in practice. There are polynomial-time algorithms, e.g., interior point methods, which are often slower in practice.

Standard problems with inequality constraints

Quadratic Programming (QP)

$$\begin{aligned} \text{minimize}_{\boldsymbol{\theta}} \quad & \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q}\boldsymbol{\theta} + \boldsymbol{c}^T\boldsymbol{\theta} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{b} \leq 0 \end{aligned}$$



If $\boldsymbol{Q}$ is positive semidefinite $\rightarrow$ convex.

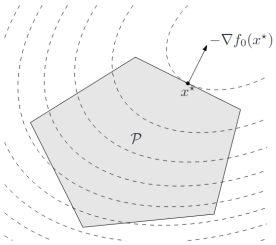
Non-negative least squares is an example of QP, with

$$\boldsymbol{\theta} = \boldsymbol{w}, \quad \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{X}^T, \quad \boldsymbol{c} = \boldsymbol{X}\boldsymbol{y}, \quad \boldsymbol{A} = -\boldsymbol{I}_D, \quad \boldsymbol{b} = 0$$

Standard problems with inequality constraints

Quadratic Programming (QP)

$$\begin{aligned} \text{minimize}_{\boldsymbol{\theta}} \quad & \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q}\boldsymbol{\theta} + \boldsymbol{c}^T\boldsymbol{\theta} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{b} \leq 0 \end{aligned}$$



If $\boldsymbol{Q}$ is positive semidefinite $\rightarrow$ convex.

Non-negative least squares is an example of QP, with

$$\boldsymbol{\theta} = \boldsymbol{w}, \quad \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{X}^T, \quad \boldsymbol{c} = \boldsymbol{X}\boldsymbol{y}, \quad \boldsymbol{A} = -\boldsymbol{I}_D, \quad \boldsymbol{b} = 0$$

We also have: Second-order cone programming, Geometric programming, ...

(Mixed) Integer Linear Programming

Integer Linear Programming (ILP)

$$\begin{aligned} \text{minimize}_{\boldsymbol{\theta}} \quad & \boldsymbol{c}^T\boldsymbol{\theta} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{b} \leq 0 \\ & \boldsymbol{\theta} \in \mathbb{Z}^D \end{aligned}$$

We now restrict the parameters $\boldsymbol{\theta}$ to be integer ($\mathbb{Z}$), or often just binary.

When a subset of the variables is real-valued, it is called a mixed ILP (or MILP).

MILPS have many applications but are NP-hard to solve in general (unlike LPs).

Outline

Motivation for projected gradient descent

Most of the time we use gradient descent to solve optimization problems.

Can we adapt gradient descent to solve constrained problems?

Constrained optimization with gradient descent

Why can't we just directly apply gradient descent for constrained optimization?

Assume parameters are restricted to convex set $\mathcal{X}$ (the domain). After some gradient descent steps, $\boldsymbol{\theta}^{t+1}$ might be outside of region $\mathcal{X}$.

Even stronger: if current point $\boldsymbol{\theta}^t$ is at the "border" of $\mathcal{X}$, each move along the gradient might lead to an infeasible solution:

$$\boldsymbol{\theta}^t \in \mathcal{X}, \text{ but } \boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \tau \nabla f(\boldsymbol{\theta}^t) \notin \mathcal{X} \text{ for all } \tau > 0$$

Constrained optimization with gradient descent

Why can't we just directly apply gradient descent for constrained optimization?

Assume parameters are restricted to convex set $\mathcal{X}$ (the domain). After some gradient descent steps, $\boldsymbol{\theta}^{t+1}$ might be outside of region $\mathcal{X}$.

Even stronger: if current point $\boldsymbol{\theta}^t$ is at the "border" of $\mathcal{X}$, each move along the gradient might lead to an infeasible solution:

$$\boldsymbol{\theta}^t \in \mathcal{X}, \text{ but } \boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \tau \nabla f(\boldsymbol{\theta}^t) \notin \mathcal{X} \text{ for all } \tau > 0$$

Idea: project a new point back to (the closest point in) the convex set $\mathcal{X}$

$$\boldsymbol{\theta}^{t+1} \leftarrow \pi_{\mathcal{X}}(\boldsymbol{\theta}^t - \tau \nabla f(\boldsymbol{\theta}^t))$$

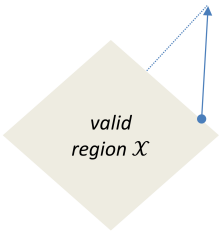
where $\pi_{\mathcal{X}}(\boldsymbol{p}) = \arg\min_{\boldsymbol{\theta} \in \mathcal{X}} \|\boldsymbol{\theta} - \boldsymbol{p}\|_2^2$ is the projection.

Projected gradient descent

Idea: Project a new point back on the convex set $\mathcal{X}$

$$\boldsymbol{\theta}^{t+1} \leftarrow \pi_{\mathcal{X}}(\boldsymbol{\theta}^t - \tau \nabla f(\boldsymbol{\theta}^t))$$

where $\pi_{\mathcal{X}}(\boldsymbol{p}) = \arg\min_{\boldsymbol{\theta} \in \mathcal{X}} \|\boldsymbol{\theta} - \boldsymbol{p}\|_2^2$.



But projection itself is a convex optimization problem!

If all constraints are linear and L_2 norm, we have a quadratic program.

- Possible to use standard solvers
- Still costly since projection has to be done after each gradient descent step

Efficient projections

Goal: $\pi_{\mathcal{X}}(\boldsymbol{p}) = \arg\min_{\boldsymbol{\theta} \in \mathcal{X}} \|\boldsymbol{\theta} - \boldsymbol{p}\|_2^2$.

Some (frequently observed) cases can be solved efficiently.

Projection onto box $\mathcal{X} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,:\, l_i \leq \theta_i \leq u_i \text{ for all } i = 1, \ldots, d\right\}$

$$\left(\pi_{\mathcal{X}}(\boldsymbol{p})\right)_i = \min(\max(l_i, p_i), u_i)$$

Projection onto L_2-ball $\mathcal{X} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,:\, \|\boldsymbol{\theta}\|_2 \leq c\right\}$

$$\pi_{\mathcal{X}}(\boldsymbol{p}) = \begin{cases} \boldsymbol{p} & \text{if } \|\boldsymbol{p}\|_2 \leq c \\ \frac{c}{\|\boldsymbol{p}\|_2}\boldsymbol{p} & \text{otherwise} \end{cases}$$

Efficient projections

Goal: $\pi_{\mathcal{X}}(\boldsymbol{p}) = \arg\min_{\boldsymbol{\theta}\in\mathcal{X}} \|\boldsymbol{\theta} - \boldsymbol{p}\|_2^2$.

Some (frequently observed) cases can be solved efficiently.

Projection onto L_1-ball $\mathcal{X} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,:\, \|\boldsymbol{\theta}\|_1 \leq c\right\}$.[3]

Projection onto L_1-ball with box-constraints
$\mathcal{X} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,:\, \|\boldsymbol{\theta}\|_1 \leq c \text{ and } l_i \leq \theta_i \leq u_i \text{ for all } i = 1, \ldots, d\right\}$.[4]

[3] Linear time algorithms: "Projection onto an L_1-norm Ball with Application to Identification of Sparse Autoregressive Models" by Jitkomut Songsiri.

[4] Linear time algorithms: "L_1 Projections with Box Constraints" by Mithun Das Gupta et al.

Projected gradient descent

Often used for solving (large-scale) constrained optimization problems.

Highly efficient if projection can be evaluated efficiently.

Each step leads to a feasible solution.[5]

Like GD, we need to choose the learning rate, etc.

Projected gradient descent is a special case of so called proximal methods.

[5] At each step, the solution remains inside the valid domain. There also exist methods that step outside the feasible region.

Outline

Constrained optimization problem

Given $f_0 : \mathbb{R}^d \to \mathbb{R}$ and $f_i : \mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}} && f_0(\boldsymbol{\theta}) \\ &\text{subject to} && f_i(\boldsymbol{\theta}) \leq 0 \quad \text{for } i = 1, \ldots, M\,. \end{aligned}$$

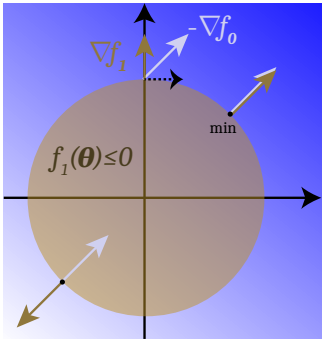
Minimization with a single inequality constraint

$$\text{minimize}_{\boldsymbol{\theta}} \quad \underbrace{-(\theta_1+\theta_2)}_{=f_0(\boldsymbol{\theta})} \qquad \text{s.t.} \qquad \underbrace{\theta_1^2+\theta_2^2-1}_{=f_1(\boldsymbol{\theta})} \leq 0$$

a) If minimizer $\boldsymbol{\theta}^*$ is in the interior of the circle $(f_1(\boldsymbol{\theta}^*) < 0)$ then $\nabla f_0(\boldsymbol{\theta}^*) = 0$.

b) If $\boldsymbol{\theta}^*$ is on the circle $(f_1(\boldsymbol{\theta}^*) = 0)$ then $-\nabla f_0(\boldsymbol{\theta}^*)$ and $\nabla f_1(\boldsymbol{\theta}^*)$ are collinear.

c) Otherwise, $-\nabla f_0(\boldsymbol{\theta}^*) = \alpha \nabla f_1(\boldsymbol{\theta}^*) + \beta \boldsymbol{v}$ for a tangent $\boldsymbol{v}$ with $\boldsymbol{v}^T \nabla f_1(\boldsymbol{\theta}^*) = 0$.



$f_0(\boldsymbol{\theta})$ is color coded.

If c) we can move[6] in direction of $\boldsymbol{v}$ to improve f_0 while still satisfying f_1.

[6]
$$f_0(\boldsymbol{\theta}^* + \epsilon \boldsymbol{v}) \simeq f_0(\boldsymbol{\theta}^*) + \epsilon \boldsymbol{v}^T \nabla f_0(\boldsymbol{\theta}^*) = f_0(\boldsymbol{\theta}^*) - \epsilon\beta \|\boldsymbol{v}\|_2^2 < f_0(\boldsymbol{\theta}^*)$$
$$f_1(\boldsymbol{\theta}^* + \epsilon \boldsymbol{v}) \simeq f_1(\boldsymbol{\theta}^*) + \epsilon \boldsymbol{v}^T \nabla f_1(\boldsymbol{\theta}^*) = f_1(\boldsymbol{\theta}^*) = 0 \rightarrow \text{ feasible!}$$

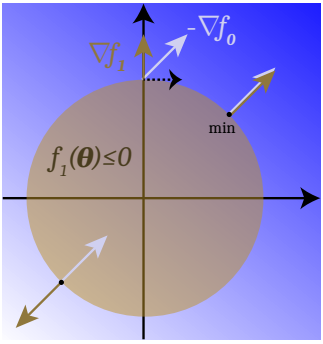
Minimization with a single inequality constraint

$$\text{minimize}_{\boldsymbol{\theta}} \quad \underbrace{-(\theta_1+\theta_2)}_{=f_0(\boldsymbol{\theta})} \qquad \text{s.t.} \qquad \underbrace{\theta_1^2+\theta_2^2-1}_{=f_1(\boldsymbol{\theta})} \leq 0$$

Thus at the minimizer $\boldsymbol{\theta}^*$ we have

$$\boxed{-\nabla f_0(\boldsymbol{\theta}^*) = \alpha\, \nabla f_1(\boldsymbol{\theta}^*)}$$

with $\alpha \geq 0$ to ensure that $-\nabla f_0$ and ∇f_1 do not point in opposite directions.



$f_0(\boldsymbol{\theta})$ is color coded.

Multiple inequality constraints

Given $f_0 : \mathbb{R}^d \to \mathbb{R}$ and $f_i : \mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}} \quad f_0(\boldsymbol{\theta}) \\ &\text{subject to} \quad f_i(\boldsymbol{\theta}) \leq 0 \quad \text{for } i = 1, \ldots, M\,. \end{aligned}$$

For multiple constraints, we have reached the minimum when the negative gradient $-\nabla f_0$ has no component $\boldsymbol{v}$ that is tangent to the boundary of the admissible set, that is $\boldsymbol{v}^T \nabla f_i(\boldsymbol{\theta}^*) = 0$ for $i = 1, \ldots, M$.

This is the case when $-\nabla f_0$ is a linear combination of the ∇f_i's with non-negative coefficients

$$\boxed{-\nabla f_0(\boldsymbol{\theta}^*) = \sum_{i=1}^{M} \alpha_i \, \nabla f_i(\boldsymbol{\theta}^*)\,, \quad \alpha_i \geq 0} \tag{1}$$

Lagrangian

$$\begin{aligned} \text{minimize}_{\boldsymbol{\theta}} \quad & f_0(\boldsymbol{\theta}) \\ \text{subject to} \quad & f_i(\boldsymbol{\theta}) \leq 0 \quad \text{for } i = 1, \ldots, M \end{aligned}$$

We define the Lagrangian $L : \mathbb{R}^d \times \mathbb{R}^M \to \mathbb{R}$ associated with the above problem as

$$\boxed{L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = f_0(\boldsymbol{\theta}) + \sum_{i=1}^{M} \alpha_i f_i(\boldsymbol{\theta})}$$

where $\alpha_i \geq 0$ is the Lagrange multiplier associated with the constraint $f_i(\boldsymbol{\theta}) \leq 0$.

Setting $\nabla_{\boldsymbol{\theta}^*} L(\boldsymbol{\theta}^*, \boldsymbol{\alpha}) = 0$ recovers the optimality criterion (Eq. 1) for $\boldsymbol{\theta}^*$,

$$\nabla_{\boldsymbol{\theta}^*} L(\boldsymbol{\theta}^*, \boldsymbol{\alpha}) = \nabla f_0(\boldsymbol{\theta}^*) + \sum_{i=1}^{M} \alpha_i \nabla f_i(\boldsymbol{\theta}^*) = 0\,.$$

Lagrange dual function

The Lagrange dual function $g : \mathbb{R}^M \to \mathbb{R}$ maps $\boldsymbol{\alpha}$ to the minimum of the Lagrangian over $\boldsymbol{\theta}$ (possibly $-\infty$ for some values of $\boldsymbol{\alpha}$),

$$g(\boldsymbol{\alpha}) = \min_{\boldsymbol{\theta} \in \mathbb{R}^d} L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left(f_0(\boldsymbol{\theta}) + \sum_{i=1}^{M} \alpha_i f_i(\boldsymbol{\theta}) \right).$$

It is concave in $\boldsymbol{\alpha}$ since it is the point-wise minimum of a family of affine functions of $\boldsymbol{\alpha}$.[7]

[7] See Boyd p. 216.

Interpretation of the Lagrangian

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}} \quad f_0(\boldsymbol{\theta}) \\ &\text{subject to} \quad f_i(\boldsymbol{\theta}) \leq 0, \quad i = 1, \ldots, M \end{aligned}$$

$$\boxed{L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = f_0(\boldsymbol{\theta}) + \sum_{i=1}^{M} \alpha_i f_i(\boldsymbol{\theta})}$$

For every choice of $\boldsymbol{\alpha}$, the corresponding *unconstrained* $g(\boldsymbol{\alpha}) = \min_{\boldsymbol{\theta} \in \mathbb{R}^d} L(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is a lower bound on the optimal value of the constrained problem:

$$\min_{\substack{\boldsymbol{\theta} \in \mathbb{R}^d \\ f_i(\boldsymbol{\theta}) \leq 0}} f_0(\boldsymbol{\theta}) = f_0(\boldsymbol{\theta}^*) \geq f_0(\boldsymbol{\theta}^*) + \underbrace{\sum_{i=1}^{M} \underbrace{\alpha_i}_{\geq 0} \underbrace{f_i(\boldsymbol{\theta}^*)}_{\leq 0}}_{\leq 0} = L(\boldsymbol{\theta}^*, \boldsymbol{\alpha}) \geq \underbrace{\min_{\boldsymbol{\theta} \in \mathbb{R}^d} L(\boldsymbol{\theta}, \boldsymbol{\alpha})}_{g(\boldsymbol{\alpha})}$$

Hence, $f_0(\boldsymbol{\theta}^*) \geq g(\boldsymbol{\alpha})$ for $\forall \boldsymbol{\alpha}$.

Lagrange dual problem

For each $\boldsymbol{\alpha} \geq \mathbf{0}$, the Lagrange dual function $g(\boldsymbol{\alpha})$ gives us a lower bound on the optimal value p^* of the original optimization problem.

What is the best (highest) lower bound?

Lagrange dual problem

$$\begin{aligned} \text{maximize}_{\boldsymbol{\alpha}} \quad & g(\boldsymbol{\alpha}) \\ \text{subject to} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

We call $\boldsymbol{\alpha}$ feasible if and only if all $\alpha_i \geq 0$ and $L(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is bounded from below[8] for $\boldsymbol{\theta} \in \mathbb{R}^d$.

[8] Since we are maximizing g, we are not interested in dual multipliers $\boldsymbol{\alpha}$ such that $g(\boldsymbol{\alpha}) = -\infty$, so the condition $g(\boldsymbol{\alpha}) \neq -\infty$ is usually added as an additional constraint to the dual problem.

Duality

Since for all $\boldsymbol{\alpha} \geq \mathbf{0}$ it holds that $g(\boldsymbol{\alpha}) \leq p^*$ we have weak duality (always),

$$d^* \leq p^* \,.$$

The difference $p^* - d^* \geq 0$ between the solution of the original and the dual problem is called the duality gap.

Under certain conditions we have strong duality where

$$d^* = p^* \,,$$

i.e. the maximum to the Lagrange dual problem is the minimum of the original (primal) constrained optimization problem ($f_0(\boldsymbol{\theta}^*) = g(\boldsymbol{\alpha}^*)$).

Dual solution

Let $\boldsymbol{\theta}^*$ be a minimizer of the primal problem and $\boldsymbol{\alpha}^*$ a maximizer of the dual problem. If strong duality holds and $L(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is convex in $\boldsymbol{\theta}$, then

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\alpha}^*)\,.$$

Proof: Recall the proof of $g(\boldsymbol{\alpha}) \leq p^*$ from Slide 21, now set $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ and use that $g(\boldsymbol{\alpha}^*) = p^*$ is given (strong duality) $\Rightarrow$ every "$\leq$" in the chain has to be a "$=$".

$$\underbrace{\min_{\boldsymbol{\theta} \in \mathbb{R}^d} L(\boldsymbol{\theta}, \boldsymbol{\alpha}^*)}_{=g(\boldsymbol{\alpha}^*)} \leq L(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^*) = f_0(\boldsymbol{\theta}^*) + \sum_{i=1}^{M} \alpha_i^* f_i(\boldsymbol{\theta}^*) \leq f_0(\boldsymbol{\theta}^*) = \underbrace{\min_{f_i(\boldsymbol{\theta}) \leq 0} f_0(\boldsymbol{\theta})}_{=p^*}$$

Under strong duality the first inequality becomes an equality and yields the claim.

Constraint qualifications for convex problems

Consider the constrained optimization problem,

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}} && f_0(\boldsymbol{\theta}) \\ &\text{subject to} && f_i(\boldsymbol{\theta}) \leq 0 \quad \text{for } i = 1, \ldots, M\,. \end{aligned}$$

Slater's constraint qualification: The duality gap is zero (i.e. strong duality holds) if $f_0, f_1, \ldots, f_M$ are convex and there exists a feasible $\boldsymbol{\theta} \in \mathbb{R}^d$ such that for every constraint $i = 1, \ldots, M$ it holds, either

- the constraint is affine, that is $f_i(\boldsymbol{\theta}) = \boldsymbol{w}_i^T \boldsymbol{\theta} + b_i$, or
- the constraint is satisfied with "$<$", that is $f_i(\boldsymbol{\theta}) \lneq 0$
  i.e., for the non-affine constraints we require a strict inequality.

Many other theorems with different conditions that guarantee zero duality gap.

For a proof see Boyd p. 234.

Recipe for solving constrained optimization problems

The constrained optimization problem,

$$\text{minimize}_{\boldsymbol{\theta}} \quad f_0(\boldsymbol{\theta}) \qquad \text{s.t.} \qquad f_i(\boldsymbol{\theta}) \leq 0 \quad \text{for } i = 1, \ldots, M\,,$$

with all $f_0, f_1, \ldots, f_M$ convex can be approached as follows:

Recipe for solving constrained optimization problems

The constrained optimization problem,

$$\text{minimize}_{\boldsymbol{\theta}} \quad f_0(\boldsymbol{\theta}) \qquad \text{s.t.} \qquad f_i(\boldsymbol{\theta}) \leq 0 \quad \text{for } i = 1, \ldots, M\,,$$

with all $f_0, f_1, \ldots, f_M$ convex can be approached as follows:

1. Formulate the Lagrangian $L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = f_0(\boldsymbol{\theta}) + \sum_{i=1}^{M} \alpha_i f_i(\boldsymbol{\theta})\,.$
2. For each $\boldsymbol{\alpha} \geq \mathbf{0}$ obtain the dual function $g(\boldsymbol{\alpha})$ by solving $g(\boldsymbol{\alpha}) = \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\alpha})$:
   2.1 Figure out for which $\boldsymbol{\alpha}$ the objective is unbounded
   2.2 For other compute $g(\boldsymbol{\alpha})$, e.g. solve $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \mathbf{0}$ to get $\boldsymbol{\theta}^*(\boldsymbol{\alpha})$, then $g(\boldsymbol{\alpha}) = L(\boldsymbol{\theta}^*(\boldsymbol{\alpha}), \boldsymbol{\alpha})$
3. Solve the dual problem, it's maximum is also the minimum of the original problem if Slater's condition is satisfied.

$$\text{maximize}_{\boldsymbol{\alpha}} \quad g(\boldsymbol{\alpha}) \qquad \text{s.t.} \qquad \alpha_i \geq 0 \quad \text{for } i = 1, \ldots, M\,.$$

KKT conditions

Assume that we have the following optimization problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{\theta}} \quad & f_0(\boldsymbol{\theta}) \\ \text{subject to} \quad & f_i(\boldsymbol{\theta}) \leq 0 \quad \text{for } i = 1, \ldots, M\,, \end{aligned}$$

with f_0 convex and $f_1, \ldots, f_M$ convex such that strong duality holds.

KKT conditions: $\boldsymbol{\theta}^*$ and $\boldsymbol{\alpha}^*$ are the optimal solutions of the constrained optimization problem and the corresponding Lagrange dual problem if and only if they satisfy the Karush-Kuhn-Tucker (KKT) conditions for $i = 1, \ldots, M$:

$$\begin{aligned} f_i(\boldsymbol{\theta}^*) &\leq 0 && \text{primal feasibility,} \\ \alpha_i^* &\geq 0 && \text{dual feasibility,} \\ \alpha_i^* f_i(\boldsymbol{\theta}^*) &= 0 && \text{complementary slackness,} \\ \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^*) &= 0 && \boldsymbol{\theta}^* \text{ minimizes Lagrangian.} \end{aligned}$$

Summary

In PGD we project the parameters to the feasible set in each step.

Lagrange formalism reformulates constrained optimization problem into an unconstrained dual problem.

For convex objective and constraints, solving the dual problem allows to solve the primal problem (duality gap is zero) under rather weak conditions.

KKT conditions characterize the solution and sometimes allow to obtain it without minimization.

Convexity is necessary to guarantee optimality.

Reading material

Main reading

- "Probabilistic Machine Learning: An Introduction" by Murphy
  [ch. 8.5.1 – 8.5.5, 8.6.1]

Additional reading

- "Convex Optimization" by Boyd
  [ch. 5.1 – 5.3, 5.5.3]

Some slides are based on an older version by S. Günnemann. Some figures are from Bishop and Murphy.