Machine Learning

Introduction to (smooth) Optimization

Prof. Matthias Hein

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Lecture 8, 13.11.2013

- Select and construct features,
- Build a model (function class, regularizer,loss, ...),
- Find best function (minimum of empirical loss and regularizer)

Optimization problem !

The Lasso: L_2-loss with L_1-regularization. How can we find w_n ?

$$w_n = \underset{w \in \mathbb{R}^D}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{D} w_j \phi_j(X_i) \Big)^2 + \lambda \sum_{i=1}^{D} |w_i|.$$

Program for today

- Convex set and functions
- How to minimize a function (Taylor expansion, Gradient descent, Newton's method),
- General Optimization Problems (Minimization with constraints)
- Convex Optimization
- Interior Point Method

Convex sets

Definition

A set C is **convex** if for any $x_1, x_2 \in C$ and for any θ with $0 \leq \theta \leq 1$ we have

$$\theta x_1 + (1-\theta)x_2 \in C.$$

Definition

A point $z = \sum_{i=1}^{k} \theta_i x_i$ where $\sum_{i=1}^{k} \theta_i = 1$ and $\theta_i \geq 0$ is a **convex combination** of $x_1, \ldots, x_k$. The **convex hull** of a set C is defined as

$$\mathrm{conv}\, C = \Big\{ \sum_{i=1}^{k} \theta_i x_i \;\Big|\; x_1, \ldots, x_k \in C,\; \theta_i \geq 0,\; \sum_{i=1}^{k} \theta_i = 1,\; k \in \mathbb{N} \Big\}.$$

- convex combination = special weighted average of the points,
- The convex hull $\mathrm{conv}\, C$ of a set C is convex. It is the smallest convex set containing C.

Convex Sets - Illustration

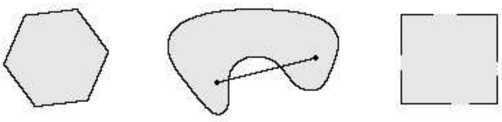


Figure : Left: Convex set, Middle: Not Convex, Right: Not Convex.

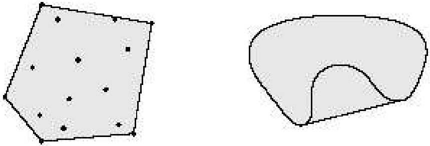


Figure : Left: convex hull of a set of points, Right: convex hull of a non-convex set.

Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex**, if

- $\operatorname{dom} f$ is a convex set,
- for all $x, y \in \operatorname{dom} f$, and $\lambda \in \mathbb{R}$ with $0 \leq \lambda \leq 1$, we have

$$f\big(\lambda x + (1-\lambda)y\big) \quad \leq \quad \lambda f(x) + (1-\lambda) f(y).$$

A function is **strictly convex** if the inequality is strict if $x \neq y$.

Further definitions:

- A function f is **concave** if and only if $-f$ is convex,
- A function f is **strictly concave** if and only if $-f$ is strictly convex,
- An affine function, $f(x) = Ax + b$, is convex and concave.

Properties of convex functions

Proposition (First order condition)

Let f be continuously differentiable and $\mathrm{dom}\, f \subseteq \mathbb{R}^n$ an open set. Then f is convex if and only if $\mathrm{dom}\, f$ is convex and

$$f(y) \geq f(x) + \langle \nabla f|_x, y - x \rangle, \quad \forall y, x \in \mathrm{dom}\, f.$$

Proposition (Second-order condition)

Let f be a twice continuously, differentiable function with open domain $\mathrm{dom}\, f$. Then f is convex if and only if $\mathrm{dom}\, f$ is a convex set and the Hessian of f, $H(f)$, is positive semi-definite for all $x \in \mathrm{dom}\, f$.

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d} \end{pmatrix}.$$

Definition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continously differentiable, then x^* is a **critical point** if

$$\nabla f(x^*) = 0 \quad \Longrightarrow \quad \text{necessary condition for a } \textbf{local minimum}.$$

Let $H(f)$ be the **Hessian** matrix

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d} \end{pmatrix}.$$

Then a sufficient condition for a local minimum at x^* is that $\nabla f(x^*) = 0$ and

$$H(f)\big|_{x^*} \succ 0 \quad \Longleftrightarrow \quad \text{for all } w \in \mathbb{R}^d \text{ with } w \neq 0, \langle w, H(f)\big|_{x^*} w \rangle > 0.$$

Taylor expansion

Key tool to derive minimization algorithms is the Taylor expansion:

Definition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be two-times continuously differentiable, then the second-order **Taylor-expansion** of the function f around x is given by

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\left\langle y - x, H(f)\big|_x (y - x)\right\rangle + o(\|y - x\|^3).$$

- **Best linear approximation** of f at x:

$$f(x) + \langle \nabla f(x), y - x\rangle .$$

- **Best quadratic approximation** of f at x:

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\left\langle y - x, H(f)\big|_x (y - x)\right\rangle .$$

General gradient descent

General gradient descent: Start with initial point x_0,

$$\text{Sequence:} \;\; x_{t+1} = x_t + \alpha_t \, d_t .$$

$x^* = \lim_{t \to \infty} x_t$ minimizes locally the function f given $\langle d_t, \nabla f(x_t) \rangle < 0$ and α_t sufficiently small (but not too small!).

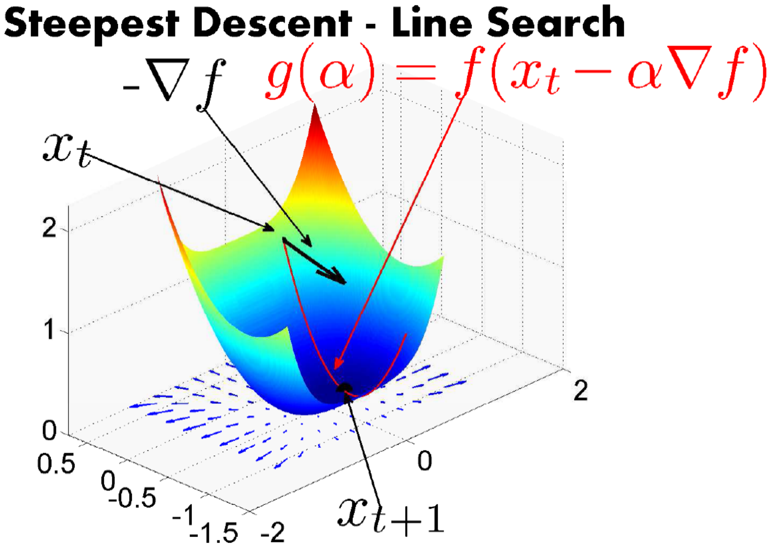
Steepest Descent:
$d_t = -\nabla f(x_t)$ (we move in the opposite direction of the gradient).

Stepsize and stopping criteria:

- α_t is the **stepsize** $\rightarrow$ has to be chosen sufficiently small, such that $f(x_{t+1}) < f(x_t)$.
  Find minimum of $g(\alpha)$ (**line search**)

  $$g(\alpha) := f(x_t + \alpha_t \, d_t)$$

- Several different **stopping criteria** e.g. $\|\nabla f(x_{t+1})\| \leq \epsilon$.


Steepest Descent - Line Search
$-\nabla f$
$g(\alpha) = f(x_t - \alpha \nabla f)$
x_t
x_{t+1}
2
1
0
0.5
0
-0.5
-1
-1.5
-2
0
2

Line Search

- Suppose g has a **single** minimum on the interval $[0, s]$,
- **Initial values:** $\alpha_0 = 0$, $\beta_0 = s$,
  Candidates: $\mu_t = \alpha_t + \tau(\beta_t - \alpha_t)$, $\qquad \nu_t = \beta_t - \tau(\beta_t - \alpha_t)$ where $\tau = \frac{3-\sqrt{5}}{2}$.
  Update rule depends on $g(\mu_t) < g(\nu_t)$, $\Longrightarrow$ always $\alpha_t \leq \beta_t$.
- The method is stopped once $\beta_{t+1} - \alpha_{t+1} < \epsilon$, where ϵ is a pre-defined threshold.
- Due to the property **golden ratio** τ,

$$\tau = (1 - \tau)^2,$$

one has to do often only one function evaluation instead of two.

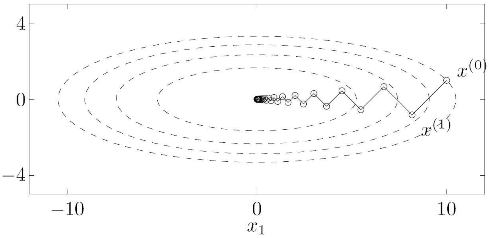
Discussion of gradient descent

Pro:

- very cheap computations
- can easily solve large-scale systems

Contra:

- sensitive to the condition number of the Hessian (ratio of largest and smallest eigenvalue) $\Longrightarrow$ elongation of level sets around local minima.
- only linear convergence $\Longrightarrow$ slow !

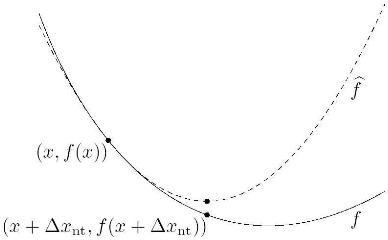


Descent direction:

$$d = -(H(f)\big|_x)^{-1}\,\nabla f(x).$$

Motivation (under assumption that Hessian is positive definite)
Descent direction d minimizes second-order approximation

$$d = \arg\min_v \hat{f}(v) = \arg\min_v \Big(f(x) + \langle \nabla f(x), v\rangle + \frac{1}{2}\left\langle v, H(f)\big|_x\, v \right\rangle \Big).$$



Gradient of $\hat{f}(v)$:

$$\nabla \hat{f} = \nabla f|_x + H(f)\big|_x v$$

Extremum at $\hat{f}(v) = 0$. Solving for v yields:

$$v = -(H(f)\big|_x)^{-1}\nabla f|_x.$$

Pro:

- superlinear convergence of Newton's method (close to the optimum),
- Newton's method is affine invariant,
- much less dependent on the choice of the parameters than gradient descent.

Contra:

- requires second derivative,
- does not scale easily to large problems if Hessian has no special structure (e.g. sparse, banded etc.) $\Longrightarrow$ one needs a fast way of solving
$$H(f)\big|_x \, d = -\nabla f.$$
- singular or non positive definite Hessian require special handling.

Summary

- Unconstrained optimization is conceptually easy
- Two standard methods:
  - steepest descent
  - Newton
- **Warning:** Only convergence to a local minimum is guaranteed !
- **General:**
  - without particular knowledge about the function convergence to global optimum is (very) difficult to achieve.
  - Global convergence can not be checked.

  Practice: Start several times with different starting vectors.
- **Convex Functions:** Every local minimum is a global minimum ! Theoretical statements about convergence rate to global minimum are possible !

Definition

A **general optimization problem** has the form

$$\min_{x \in D} f(x),$$
$$\text{subject to: } g_i(x) \leq 0, \; i = 1, \ldots, r$$
$$h_j(x) = 0, \; j = 1, \ldots, s.$$

- x is the optimization variable, f the objective (cost) function,
- $x \in D$ is **feasible** if the inequality and equality constraints hold at x.
- the **optimal value** p^* of the optimization problem

$$p^* = \inf\{ f(x) \,|\, x \text{ feasible } \}.$$

$p^* = -\infty$: problem is unbounded from below,
$p^* = \infty$: problem is infeasible.

Terminology:

- given that x is a **feasible** point,
  $g_i(x) = 0$: inequality constraint is **active** at x.
  $g_i(x) < 0$: is **inactive**.
  A constraint is **redundant** if deleting it does not change the feasible set.
- A point x is called **locally optimal** if there exists $R > 0$ such that

$$f(x) = \inf\{f(z) \mid \|z - x\| \leq R,\ z \text{ feasible } \}.$$

Definition

A **convex optimization problem** has the standard form

$$\min_{x \in D} f(x),$$
$$\text{subject to: } g_i(x) \le 0, \; i = 1, \ldots, r$$
$$\langle a_j, x \rangle = b_j, \; j = 1, \ldots, s,$$

where $f, g_1, \ldots, g_r$ are convex functions.

Difference to the general problem:

- the objective function f has to be convex (LP: linear, QP: quadratic),
- the inequality constraint functions g_i have to be convex (LP,QP: linear),
- the equality constraint function have to be linear.

$\Longrightarrow$ **The feasible set of a convex optimization problem is convex.**

Local and global minima of convex optimization problems

Theorem

- *Any locally optimal point of a convex optimization problem is globally optimal.*
- *The set of global optima is convex.*
- *If the objective function f is strictly convex, then the global optimum is unique.*

Proof: Suppose x is locally optimal, that means x is feasible and $\exists\, R > 0$,

$$f(x) = \inf\{f(z) \,|\, \|z - x\| \leq R,\ z \text{ feasible } \}.$$

Assume x is not globally optimal $\Longrightarrow \exists$ feasible y such that $f(y) < f(x)$.

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) < f(x),$$

for any $0 < \lambda < 1 \implies x$ is not locally optimal $\lightning$.

The Lagrange function

Motivation of the Lagrange function: general optimization problem

$$\begin{aligned} &\min_{x \in D} f(x), \\ \text{subject to: } & g_i(x) \leq 0, \; i = 1, \ldots, r \\ & h_j(x) = 0, \; j = 1, \ldots, s. \end{aligned}$$

Idea:

- turn constrained problem into an unconstrained problem,
- the extremal points of the unconstrained problem contain the extremal points of the constrained problem (necessary condition) and in some cases the two sets are equal (necessary and sufficient condition).

The Lagrange function II

Definition

The **Lagrangian** or **Lagrange function** $L : \mathbb{R}^n \times \mathbb{R}^r_+ \times \mathbb{R}^s \to \mathbb{R}$ associated with the MP is defined as

$$L(x, \lambda, \mu) = f(x) + \sum_{j=0}^{r} \lambda_j \, g_j(x) + \sum_{i=0}^{s} \mu_i \, h_i(x),$$

with $\operatorname{dom} L = D \times \mathbb{R}^r_+ \times \mathbb{R}^s$ where D is the domain of the optimization problem. The variables λ_j and μ_i are called **Lagrange multipliers** associated with the inequality and equality constraints.

Note: The Lagrange multipliers $\{\lambda_j\}_{j=1}^r$ of inequality constraints are non-negative !

For all feasible x,

$$\sum_{j=0}^{r} \lambda_j \, g_j(x) + \sum_{i=0}^{s} \mu_i \, h_i(x) \leq 0 \quad \Longrightarrow \quad \mathbf{L(x, \lambda, \mu) \leq f(x)}.$$

The dual function

Definition

The **dual Lagrange function** $q : \mathbb{R}^r_+ \times \mathbb{R}^s \to \mathbb{R}$ associated with the MP is defined as

$$q(\lambda, \mu) = \inf_{x \in D} L(x, \lambda, \mu) = \inf_{x \in D} \Big(f(x) + \sum_{j=0}^{r} \lambda_j \, g_j(x) + \sum_{i=0}^{s} \mu_i \, h_i(x) \Big),$$

where $q(\lambda, \mu)$ is defined to be $-\infty$ if $L(x, \lambda, \mu)$ is unbounded from below in x.

Properties:

- the dual function is a pointwise infimum of a family of concave functions (in λ and μ) and therefore concave.
  This holds irrespectively of the character of the MP, in particular this holds also for discrete optimization problems.
- For any $\lambda \succeq 0$ and $\mu \in \mathbb{R}^s$, feasible x',

$$\inf_{x \in D} L(x, \lambda, \mu) \leq f(x') \quad \Longrightarrow \quad \mathbf{q}(\lambda, \mu) \leq \mathbf{p}^*.$$

The dual problem

For each pair (λ, μ) with $\lambda \succeq 0$ we have $q(\lambda, \mu) \leq p^*$.

What is the best possible lower bound ?

Definition

The **Lagrange dual problem** is defined as

$$\max_{\lambda, \mu} q(\lambda, \mu),$$

$$\text{subject to: } \lambda_i \geq 0, \ i = 1, \ldots, r.$$

Properties:

- For each MP the dual problem is **convex**.
- The original OP is called the **primal problem**.
- (λ, μ) is **dual feasible** if $q(\lambda, \mu) > -\infty$.
- (λ^*, μ^*) is called **dual optimal** if they are optimal for the dual problem.

Weak and strong duality

Corollary

Let d^* and p^* be the optimal values of the dual and primal problem. Then

$$d^* \leq p^*, \qquad \textbf{(weak duality)}.$$

- The difference $p^* - d^*$ is the **optimal duality gap** of the MP.
- solving the convex dual problem provides lower bounds for any MP.

Definition

We say that **strong duality** holds if

$$d^* = p^*.$$

Constraint qualifications are conditions under which strong duality holds.

Strong duality **does not** hold in general !
But for convex problems strong duality holds most of the time.

Slater's constraint qualification:

Theorem

Suppose that the primal problem is convex and there exists an $x \in \mathrm{relint}\, D$ *such that*

$$g_i(x) < 0, \quad i = 1, \ldots, r,$$

then Slater's condition holds and strong duality holds. Strict inequality is not necessary if $g_i(x)$ *is an affine constraint.*

- What is an interior point x of a set D ?
  There exists a $\varepsilon > 0$ such that the ball around x of radius ε is contained in D $\Rightarrow$ a subspace has always empty interior !
- The relative interior of D is the interior relative to the subpace it is lying in.

Why is the dual problem useful ?

Measure of suboptimality

- every dual feasible point (λ, μ) provides a **certificate** that $p^* \geq q(\lambda, \mu)$,
- every feasible point x provides a **certificate** that $d^* \leq f(x)$,
- any primal/dual feasible pair x and (λ, μ) provides an upper bound on the duality gap: $f(x) - q(\lambda, \mu)$, or

$$p^* \in [q(\lambda, \mu), f(x)], \qquad d^* \in [q(\lambda, \mu), f(x)].$$

- duality gap is zero $\Longrightarrow$ x and (λ, μ) is primal/dual optimal.
- **Stopping criterion:** for an optimization algorithm which produces a sequence of primal feasible x_k and dual feasible (λ_k, μ_k). If strong duality holds use:

$$f(x_k) - q(\lambda_k, \mu_k) \leq \varepsilon.$$

KKT optimality conditions

Theorem

- f, g_i and h_j differentiable,
- strong duality holds.

Then necessary conditions for primal and dual optimal points x^* and (λ^*, μ^*) are the **Karush-Kuhn-Tucker(KKT) conditions**

$$\begin{aligned} g_i(x^*) \leq 0,\ i = 1, \ldots, r, \qquad & h_j(x^*) = 0,\ j = 1, \ldots, s, \\ \lambda_i^* \geq 0,\ i = 1, \ldots, r \qquad & \lambda_i^* g_i(x^*) = 0,\ i = 1, \ldots, r \end{aligned}$$

$$\nabla f(x^*) + \sum_{i=1}^{r} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{s} \mu_j^* \nabla h_j(x^*) = 0.$$

If the primal problem is **convex**, then the KKT conditions are **necessary and sufficient** for primal and dual optimal points with zero duality gap.

Remarks

- The condition:

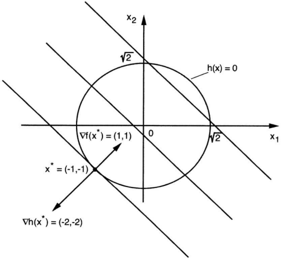
$$\nabla f(x^*) + \sum_{i=1}^{r} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{s} \mu_j^* \nabla h_j(x^*) = 0,$$

  is equivalent to $\nabla_x L(x, \lambda^*, \mu^*)\Big|_{x^*} = 0$.
- **convex problem:** any pair x, (λ, μ) which fulfills the KKT-conditions is primal and dual optimal. **Additionally:** Slater's condition holds $\Longrightarrow$ such a point exists.
- Assume: strong duality and a dual optimal solution (λ^*, μ^*) is known and $L(x, \lambda^*, \mu^*)$ has a unique minimizer x^*
  1. x^* is primal optimal as long as x^* is primal feasible,
  2. If x^* is not primal feasible, then the primal optimal solution is not attained.

Geometric Interpretation for an equality constraint:

- The set, $h_i(x) = 0$, $i = 1, \ldots, m$, determines a constraint surface in $\mathbb{R}^d$.
- First order variations of the constraints (tangent space of the constraint surface)

$$h(x) = h(x^*) + \langle \nabla h(x^*), x - x^* \rangle \approx 0 \quad \Longrightarrow \quad \langle \nabla h(x^*), x - x^* \rangle = 0.$$

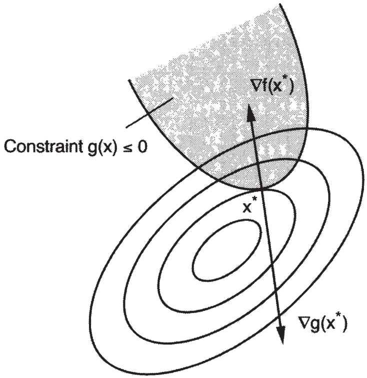


- at a local minima x^* the gradient ∇f is orthogonal to the subspace of first order variations

$$V(x^*) = \{w \in \mathbb{R}^d \mid \langle w, \nabla h_i(x^*) \rangle = 0, \ i = 1$$

- Equivalently, $\nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0.$

Geometric Interpretation for an inequality constraint:



Two cases:

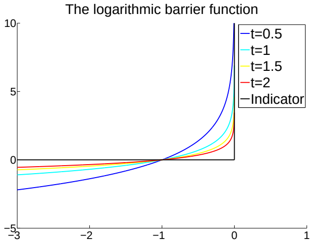
- constraint active: $g(x^*) = 0$:
$$\nabla f(x^*) + \lambda \nabla g(x^*) = 0.$$
- constraint inactive: $g(x^*) < 0$,
$$\nabla f(x^*) = 0.$$

Equivalent formulation of general convex optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x) + \sum_{i=1}^m I_-(g_i(x))$$

subject to: $Ax = b$,

where $I_-(u) = \begin{cases} 0, & u \leq 0 \\ \infty, & u > 0. \end{cases}$.



Basic idea: approximate indicator function with a differentiable function with closed level sets.

$$\hat{I}_-(u) = -\left(\frac{1}{t}\right)\log(-u), \qquad \mathrm{dom}\,\hat{I} = \{x \,|\, x < 0\}.$$

where t is a parameter controlling the accuracy of the approximation.

Definition of barrier function: $\phi(x) = -\sum_{i=1}^{m} \log(-g_i(x))$.

Approximate formulation:

$$\min_{x\in\mathbb{R}^n} \; t\, f(x) + \phi(x)$$

$$\text{subject to: } Ax = b,$$

Definition

Let $x^*(t)$ be the optimal point of the above problem, which is called **central point**. The **central path** is the set of points $\{x^*(t) \,|\, t > 0\}$.

How is the new optimization problem related to the original one ?

$$f(x^*(t)) - p^* \leq \frac{m}{t}.$$

As $t \to \infty$, we obtain the solution of the original problem !

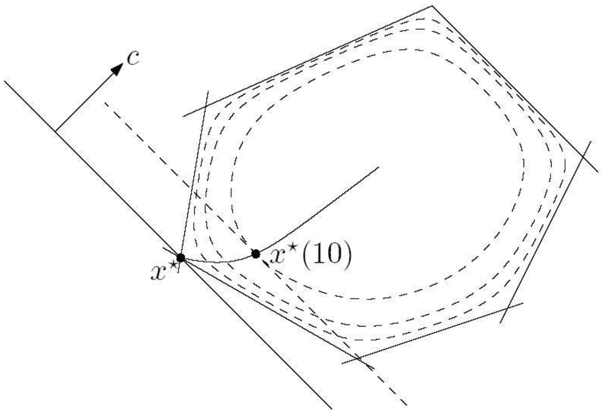


Figure : The central path for an LP. The dashed lines are the the contour lines of ϕ. The central path converges to x^* (solution of the original problem) as $t \to \infty$.

The barrier method (direct): set $t = \frac{m}{\varepsilon}$ then

$$f(x^*(t)) - p^* \leq \varepsilon \;\Rightarrow\; \textbf{numerical problems in practice.}$$

Barrier method or path-following method:

Require: strictly feasible x^0, γ, $t = t^{(0)} > 0$, tolerance $\varepsilon > 0$.

1: **repeat**

2: Centering step: compute $x^*(t)$ by minimizing

$$\min_{x \in \mathbb{R}^n} \; t\, f(x) + \phi(x)$$
$$\text{subject to: } \; Ax = b,$$

where **previous central point is taken as starting point.**

3: **UPDATE:** $x = x^*(t)$.

4: $t = \gamma t$.

5: **until** $\frac{m\gamma}{t} < \varepsilon$