# Lecture 10

Waldemar Hebisch

December 21, 2021

## 1 Duality

Recall optimization problem with constraints: given a set $S$ of form

$$S = \{x : \forall_{i \in E} g_i(x) = 0, \forall_{i \in I} g_i(x) \leq 0\}$$

where $E$ is called set of equality constraints, $I$ is called set of inequality constraints, we want to minimize $f : S \to \mathbb{R}$. We collect all constraints into single vector valued function $g$:

$$g(x) = (g_1(x), \ldots, g_m(x))$$

and motivated by KKT conditions write

$$L(x, \lambda) = f(x) + \langle \lambda, g(x) \rangle$$

where $L$ is called Lagrangian (or Lagrange function).

In terms of Lagrangian we can rewrite KKT conditions as

$$\partial_x L(x, \lambda) = 0$$

and for $i \in I$ we have $\lambda_i \geq 0$ and $\lambda_i = 0$ if $g_i(x) < 0$.

We define dual function $h$ by formula

$$h(\lambda) = \inf_x L(x, \lambda)$$

with convention that when $L$ for given $\lambda$ is unbounded from below as function of $x$, then $\lambda$ is not in domain of $h$. We would get equivalent results taking $-\infty$ as value of $h(\lambda)$. Since the dual function is the pointwise infimum of a family of affine functions of $\lambda$ it is concave even for non-convex problems. In particular domain of $h$ is convex.

Note: Above minimization is unconstrained, that is over whole domain, which usually is bigger than feasible set.

Note: Points where $f(x) = \infty$ (or $g_i(x) = \infty$) do not affect $h(\lambda)$, so it does not matter if we include then in domain of $f$ (respectively $g_i$).

## 1.1 Conjugate function

Example: Consider (trivial) problem of minimizing $f(x)$ under constraint $x = 0$. Lagrange function is

$$L(x, \lambda) = f(x) + \langle x, \lambda \rangle$$

and

$$h(\lambda) = \inf_x (f(x) + \langle x, \lambda \rangle) = -\sup(\langle x, -\lambda \rangle - f(x)).$$

The function

$$f^*(\lambda) = \sup_x (\langle x, \lambda \rangle - f(x))$$

is called Legendre transform or conjugate of $f$. With this notation we have

$$h(\lambda) = -f^*(-\lambda)$$

To get more explicit example, consider single variable $f(x) = x \log(x)$. To maximize $\langle x, \lambda \rangle - f(x)$ compute derivative

$$\partial_x (\langle x, \lambda \rangle - f(x)) = \lambda - \log(x) - 1.$$

Hence, for optimal $x$ we have

$$\log(x) = \lambda - 1$$

that is

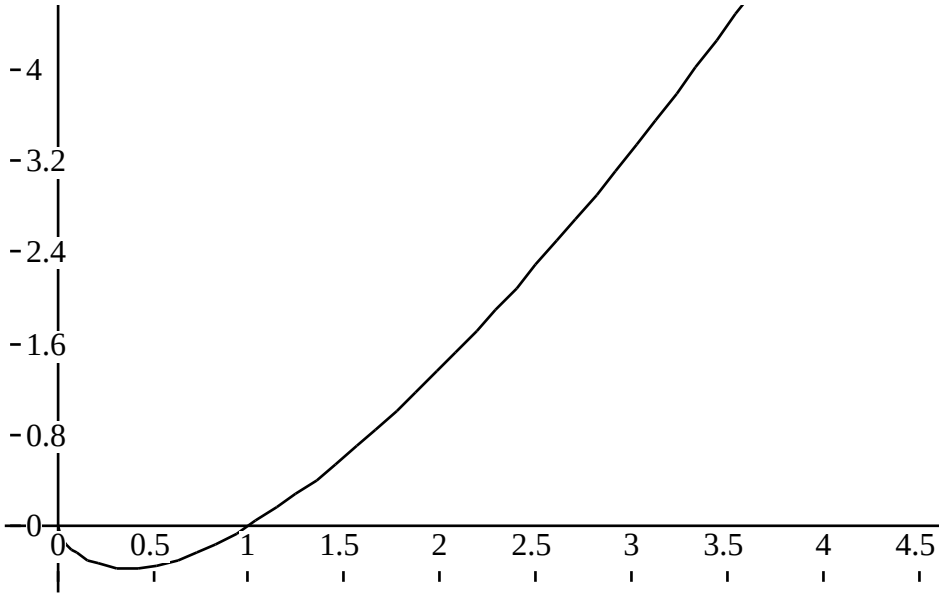
$$x = \exp(\lambda - 1).$$

Now

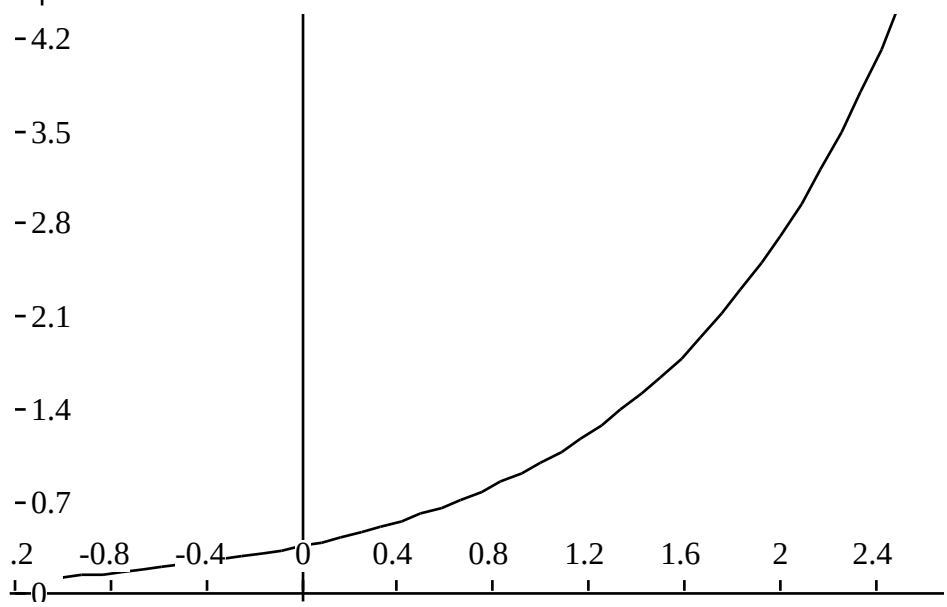
$$\begin{aligned} \langle x, \lambda \rangle - f(x) &= \exp(\lambda - 1)\lambda - \exp(\lambda - 1) \log(\exp(\lambda - 1)) \\ &= \exp(\lambda - 1)\lambda - \exp(\lambda - 1)(\lambda - 1) = \exp(\lambda - 1) \end{aligned}$$

so

$$f^*(\lambda) = \exp(\lambda - 1)$$




Graph of $x\log(x)$.




Graph of $\exp(\lambda-1)$.

From the above we see that

$$h(x) = -f^*(-\lambda) = -\exp(-\lambda - 1).$$

Similar, calculation shows that for $p > 1$ and $f(x) = \frac{1}{p}x^p$ we have

$$f^*(\lambda) = \frac{1}{q}\lambda^q$$

where $q$ solves $\frac{1}{p} + \frac{1}{q} = 1$, that is $q = \frac{p}{p-1}$.

For $f(x) = |x|$, when $|\lambda| > 1$ the expression

$$\langle x, \lambda \rangle - f(x) = x\lambda - |x|$$

is unbounded, so $\lambda$ is not in domain of $f^*$. When $\lambda \in [-1, 1]$, then expression above is non-positive and maximum is attained for $x = 0$, so $f^*(\lambda) = 0$.

For $f(x) = 0$ with domain $[-1, 1]$, we get

$$f^*(\lambda) = \sup_{x \in [-1,1]} x\lambda.$$

When $\lambda \geq 0$ maximum is at $x = 1$, so $f^*(\lambda) = \lambda$. When $\lambda \leq 0$ maximum is at $x = -1$ so $f^*(\lambda) = -\lambda$. In other words

$$f^*(\lambda) = |\lambda|.$$

Recalling previous example, we get

$$(f^*)^* = f.$$

This in fact is general equality for convex $f$.

Let $B = \{x : \|x\| \leq 1\}$ be unit ball in $\mathbb{R}^n$ and let $f(x) = 0$ with domain $B$. We have

$$f^*(\lambda) = \sup_{x \in B} \langle x, \lambda \rangle$$

Clearly,

$$\langle x, \lambda \rangle \leq \|x\| \|\lambda\| \leq \|\lambda\|.$$

For nonzero $\lambda$ taking $x = \frac{\lambda}{\|\lambda\|}$ we have

$$\langle x, \lambda \rangle = \langle \frac{\lambda}{\|\lambda\|}, \lambda \rangle = \frac{\|\lambda\|^2}{\|\lambda\|} = \|\lambda\|$$

so

$$f^*(\lambda) = \|\lambda\|$$

(it is easy to check that this equality is also true for $\lambda = 0$).

Remark: There is similar equality for balls corresponding to $l^p$ norm with $p > 1$, however in such case $f^*$ contains $l^q$ norm, where $q$ solves equality $\frac{1}{p} + \frac{1}{q} = 1$. This can be generalized further, for ball corresponding to given norm we get dual norm.

For quadratic $f(x) = \frac{1}{2}\langle Ax, x \rangle$ with strictly positive definite $A$ we have

$$f^*(\lambda) = \frac{1}{2}\langle A^{-1}\lambda, \lambda \rangle.$$

When $A$ is invertible matrix $u(x) = f(Ax)$, then

$$u^*(\lambda) = f^*((A^T)^{-1}\lambda).$$

## 1.2 Examples of dual problems

Example: For linear programming problem in standard form, that is $f(x) = \langle c, x\rangle$ with constraints $Ax + b = 0$, $x_i \geq 0$ it is useful to write $\lambda$ as pair $(\eta, \theta)$ where $\eta$ corresponds to equality constraints and $\theta$ corresponds to inequality constraints. Then (using $g(x) = -x$)

$$L(x, \eta, \theta) = \langle c, x\rangle + \langle \eta, Ax + b\rangle - \langle \theta, x\rangle$$

$$= \langle c, x\rangle + \langle \eta, b\rangle + \langle A^T\eta, x\rangle - \langle \theta, x\rangle$$

$$= \langle \eta, b\rangle + \langle c + A^T\eta - \theta, x\rangle$$

so dual $h$ is

$$h(\eta, \theta) = \inf_x \langle \eta, b\rangle + \langle c + A^T\eta - \theta, x\rangle.$$

Linear function is bounded from below only when it identically 0.

So we get

$$h(\eta, \theta) = \langle \eta, b\rangle$$

with domain consisting of pairs $(\eta, \theta)$ such that

$$c + A^T\eta - \theta = 0.$$

Example: Linear programming problem in inequality form, that is $f(x) = \langle c, x\rangle$ with constraints $Ax + b \geq 0$. This time

$$L(x, \lambda) = \langle c, x\rangle - \langle \lambda, Ax + b\rangle$$

and calculation as for standard form gives

$$h(\lambda) = -\langle \lambda, b\rangle$$

with domain consisting of $\lambda$ such that $c - A^T\lambda = 0$.

Example: quadratic $f(x) = \frac{1}{2}\langle x, x\rangle$ with constraints $Ax + b = 0$. Then

$$L(x, \lambda) = \frac{1}{2}\langle x, x\rangle + \langle \lambda, Ax + b\rangle$$

which is strictly convex quadratic function. We minimize it looking for zero of derivative with respect to $x$:

$$x + A^T\lambda = 0$$

which gives

$$h(\lambda) = L(-A^T\lambda, \lambda) = -\frac{1}{2}\langle AA^T\lambda, \lambda\rangle + \langle \lambda, b\rangle.$$

## 1.3 Weak and strong duality

**Lemma 1.1** *When $\lambda_i \geq 0$ for $i \in I$, then*

$$h(\lambda) \leq \inf_{x \in S} f(x)$$

Remark: this is called weak duality.
Proof: If $y$ is feasible point, then

$$L(y, \lambda) = f(y) + \sum_i \lambda_i g_i(y) \leq f(y).$$

Namely, when $i \in E$ and $y \in S$, then $g_i(y) = 0$ so

$$\sum_{i \in E} \lambda_i g_i(y) = 0.$$

When $i \in I$ and $y \in S$, then $g_i(y) \leq 0$ and $\lambda_i \geq 0$ so

$$\sum_{i \in I} \lambda_i g_i(y) \leq 0$$

which gives inequality above.
Now,

$$h(\lambda) = \inf_x L(x, \lambda) \leq L(y, \lambda) \leq f(y)$$

so indeed

$$h(\lambda) \leq \inf_{x \in S} f(x).$$

Given lemma about weak duality it is natural to seek best possible lower bound, that is put

$$C = \{\lambda : \forall_{i \in I} \lambda_i \geq 0\}$$

and maximize $h(\lambda)$ over $C$. This is called dual problem. Dual problem is convex even if original problem (usually called primal problem) is non-convex. It is interesting to ask when bound obtained from dual problem is tight.
In general we call difference

$$\inf_{x \in S} f(x) - \sup_{\lambda \in C} h(\lambda)$$

duality gap. When duality gap is zero we say that strong duality holds. Duality gap may be positive. But in important special cases strong duality holds.

**Lemma 1.2** *If $f$ and $g_i$ are convex, and KKT conditions hold for $x$ and $\lambda$, then $x$ is optimal solution to primal problem, $\lambda$ is optimal solution to dual problem and*

$$h(\lambda) = f(x)$$

*that is strong duality holds.*

Proof: Since $f$ and $g_i$ are convex Lagrangian $L$ is convex as function of $x$. KKT conditions mean that

$$\partial_x L(x, \lambda) = 0.$$

Since $L$ is convex with respect to $x$, this means that

$$L(x, \lambda) = \inf_y L(y, \lambda) = h(\lambda).$$

However, in KKT conditions $g_i(x) = 0$ when $\lambda_i \neq 0$, so $\lambda_i g_i(x) = 0$ and

$$L(x, \lambda) = f(x) + \sum \lambda_i g_i(x) = f(x)$$

so

$$f(x) = h(\lambda).$$

Since $h(\lambda)$ is lower bound on optimal solution, $f(x)$ is optimal value for primal problem and $x$ is solution to primal problem. Similarly $\lambda$ is solution to dual problem.

Remark: In general, when $f(x) = h(\lambda)$, then $x$ is optimal solution to primal problem and $\lambda$ is optimal solution to dual problem.

**Lemma 1.3** *When $f$ is convex differentiable and all $g_i$ are affine, then KKT conditions for optimal solution and consequently strong duality hold. In particular strong duality holds for linear programming problems and quadratic problems.*

Proof: Recall tangent cone $TS_x$. $v \in TS_x$ if and only if

$$\lim_{t \to 0_+} \frac{d(x + tv, S)}{t} = 0.$$

We know that if $v \in TS_x$ for active inequality constraints we have

$$\langle \nabla g_i(x), v \rangle \leq 0.$$

But for affine $g_i$ we have

$$g_i(x + tv) = g_i(x) + t\langle \nabla g_i(x), v \rangle$$

so $g_i(x + tv) \leq 0$.

For equality constraints we have $\langle \nabla g_i(x), v \rangle = 0$ and similar to above $g_i(x + tv) = 0$.

For inactive constraints when $t$ is small enough we have $g_i(x + tv) \leq 0$.

Together, when $\langle \nabla g_i(x), v \rangle = 0$ for equality constraints, and $\langle \nabla g_i(x), v \rangle \leq 0$ for all active inequality constraints, then $x + tv \in S$ for small enough $t \geq 0$, that is $v \in TS_x$. In other words

$$TS_x = \{v : \forall_{i \in E} \langle \nabla g_i(x), v \rangle = 0, \quad \forall_{i \in J} \langle \nabla g_i(x), v \rangle \leq 0\}$$

where $J$ is set of active inequality constraints at $x$.

But this equality was all that we needed to obtain KKT conditions.

Example: For linear programming problem: minimize $\langle c, x\rangle$ with constraints $Ax + b = 0$, $x \geq 0$ we get as dual problem: maximize $(\eta, b)$ with domain domain consisting of pairs $(\eta, \theta)$ such that

$$c + A^T\eta + \theta = 0.$$

under constraint $\theta \geq 0$. Formally this is not a linear programming problem, but we can treat equality $A^T\eta + \theta + c = 0$ as constraint, so dual problem is equivalent to linear programming problem. Note that conditions $\theta \geq 0$ and $A^T\eta + \theta + c = 0$ are equivalent to $A^T\eta + c \leq 0$. So we can write dual problem as

$$\begin{array}{ll} \text{maximize} & \langle \eta, b\rangle \\ \text{subject to} & A^T\eta + c \leq 0 \end{array}$$

In linear programming solution solution to dual problem gives optimal value for primal problem. Moreover, from solution to dual problem we get set of active constraints which gives basic set corresponding to optimal solution of primal problem. Knowing basic set we can find optimal solution of primal problem by solving linear equations. When dual problem is solved using simplex method such approach is called dual simplex method.

## 1.4 Interior point methods

Simplex method solves linear programming problem by moving trough boundary points of feasible set. We already mentioned that interior point methods find sequence of point in the interior of feasible set that converges to optimal solution. In particular we will look at barrier methods (which are now dominant form of interior point methods). For linear constraints it is usual to use logarithmic barrier. More precisely, we replace inequality constraints $x \geq 0$ by barrier

$$\phi(x) = -\sum_{i=1}^{m} \log(x_i)$$

Equality constraints remain (and must be treated separately). For $\lambda > 0$ we get problem of minimizing

$$f(x) + \lambda\phi(x)$$

This is equivalent to minimizing

$$\frac{1}{\lambda} f(x) + \phi(x)$$

Since $f$ is affine (as we have linear programming problem), function above is self-concordant, so we have good convergence properties for Newton method. Changing a bit notation we write

$$u_\theta(x) = \theta f(x) + \phi(x)$$

and consider problem of minimizing $u_\theta$ under equality constraints $Ax + b = 0$. For $\theta \geq 0$ denote by $x_\theta$ optimal solution of the problem above (we assume that solution exists). $x_\theta$ is called central path.

When $\theta$ goes to infinity, then under mild regularity conditions $x_\theta$ goes to $x_\infty$, where $x_\infty$ is optimal solution of original problem.

For $u_\theta$ Lagrangian is

$$L_\theta(x, \lambda) = \theta f(x) + \phi(x) + \langle Ax + b, \lambda \rangle$$

$$= \theta f(x) + \phi(x) + \langle b, \lambda \rangle + \langle A^T \lambda, x \rangle$$

so KKT conditions are

$$\theta \nabla f(x_\theta) + \nabla \phi(x_\theta) + A^T \lambda = 0$$

Now put $\psi(x) = \langle \nabla \phi(x_\theta), x \rangle$. By equality above KKT conditions at $x_\theta$ are satisfied for $\theta f(x) + \psi(x)$ so

$$\theta f(x_\infty) + \psi(x_\infty) \geq \theta f(x_\theta) + \psi(x_\theta)$$

But coordinates of $\nabla \phi(x)$ are negative so $\psi(x) \leq 0$.

Since $(\nabla \phi(x))_i = \frac{-1}{x_i}$ we have

$$\psi(x_\theta) = \langle \nabla \phi(x_\theta), x_\theta \rangle = -n$$

so

$$\theta f(x_\infty) \geq \theta f(x_\theta) - n$$

that is

$$f(x_\theta) - f(x_\infty) \leq \frac{n}{\theta}$$

so we can easily estimate how far $f(x_\theta)$ is from optimal value. In particular, to be at most $\varepsilon$ from optimal value we need $\theta \geq \frac{n}{\varepsilon}$.

Remark: Our use of KKT conditions in deriving estimate above is equivalent to using feasible point of dual problem, but is much more direct.

Our error estimate suggest that we could take $\theta \geq \frac{n}{\varepsilon}$ and use Newton method on equality constrained problem to find $x_\theta$. But far from optimal point Newton method may be relatively slow.

So better approach is to start from relatively small $\theta$. For such $\theta$ corresponding $x_\theta$ is well inside interior of feasible set and we can expect reasonably fast convergence. Then we successively multiply $\theta$ by a constant and solve new problem using approximate solution of previous one as a starting point. In pseudocode, using $x_0$ as initial approximation:

1. take $i = 0$ and $\theta_0 = 1$
2. approximately minimize $u_{\theta_i} = \theta_i f(x) + \phi(x)$ using $x_i$ as starting point. Call the result $x_{i+1}$
3. put $\theta_{i+1} = a\theta_i$

4. increase $i$ by 1 and go to step 2

It remains to choose parameter $a$. Theoretically best result are obtained when $a-1$ is small fraction of $\frac{1}{\sqrt{n}}$. We can explain idea in case when there are no equality constraints. Newton method depends on $\nabla^2 u_\theta$. Since our $f$ is affine second derivative is independent of $\theta$ and we get $\nabla^2 u_\theta = \nabla^2 \phi$ and we have formula

$$(\nabla^2 \phi(x))_{i,i} = \frac{1}{x_i^2}$$

while elements outside diagonal are 0. Due to KKT conditions

$$\theta \nabla f(x_\theta) + \nabla \phi(x_\theta) = 0$$

so

$$\theta \nabla f(x_\theta) = -\nabla \phi(x_\theta)$$

In terms of coordinates

$$\theta (\nabla f(x_\theta))_i = \frac{1}{(x_\theta)_i}$$

which means that

$$\langle (\nabla^2 \phi(x_\theta))^{-1} \theta \nabla f(x_\theta), \theta \nabla f(x_\theta) \rangle = n$$

Now, due to KKT conditions

$$\nabla u_{a\theta}(x_\theta) = (a-1)\theta \nabla f(x_\theta).$$

Recall that for convergence of Newton method applied to $f$ critical quantity is

$$\lambda(f,x) = \langle (\nabla^2 f(x))^{-1} \nabla f(x), \nabla f(x) \rangle^{1/2}$$

And we had estimate after single step:

$$\lambda(f, x_{i+1}) \leq 2\lambda(f, x_i)^2.$$

In our case $\lambda(u_{a\theta}, x_\theta)$ is

$$\lambda(u_{a\theta}, x_\theta) = \langle (\nabla^2 u_{a\theta}(x_\theta))^{-1} \nabla u_{a\theta}(x_\theta), \nabla u_{a\theta}(x_\theta) \rangle^{1/2}$$

$$= |a-1| \langle (\nabla \phi)(x_\theta))^{-1} \theta \nabla f(x_\theta), \theta \nabla f(x_\theta) \rangle^{1/2} = |a-1|\sqrt{n}$$

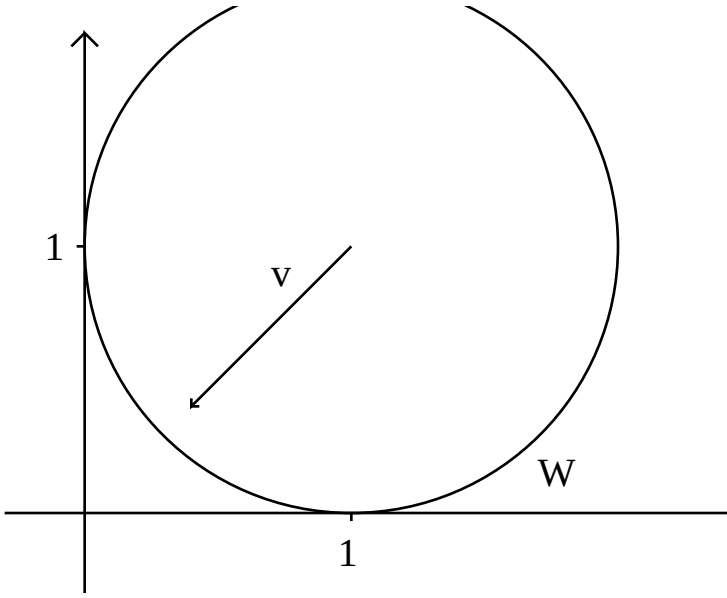
So, when $|a-1|\sqrt{n} \leq \frac{1}{4}$ then one step of damped Newton method at least halves $\lambda(u_{a\theta}, x)$, so single step on damped Newton method per increase of $\theta$ is enough. Rigorous proof must take into account that we start each step only at approximate $x_\theta$ and handle constraints, but this can be done with slightly smaller $a$.

So one can prove that $O(\sqrt{n}\log(\frac{1}{\varepsilon}))$ iterations is enough to get accuracy $\varepsilon$. This is somewhat pessimistic and in practice using $a$ between 2 and 10 one gets much faster convergence.

Intuitively, this can be understood forgetting about equality constraints. Since Newton method is affine invariant we can rescale coordinates so that $x_\theta = (1, \ldots, 1)$. Note that rescaling only changes $\phi$ by additive constant. Then due to KKT condition and form of $\phi$ we have

$$\nabla f(x) = (1, \ldots, 1).$$

Also, 0 is optimal point so for fast convergence we need $\nabla u_{a\theta}(x)$ to be comparable to $\nabla f(x)$. In other words we should rather double $\theta$. And with proper choice of step size bigger $a$ is better. Alas, with constraints we may be forced to move in somewhat different direction and it is not clear if argument above can be made rigorous.




Geometrically we have the following picture:

Self-concordant estimate works well only on ball $W$. $v$ gives direction of derivative for Newton method. Clearly we would like to move as close to origin as possible (origin is the optimal point). However distance to origin is larger than radius of $W$. On the plane we get factor of $\sqrt{2}$ which does not look bad. But in high dimension distance to origin is $\sqrt{n}$ times bigger than radius of $W$ and this makes significant difference. The $a$ parameter implicitly decides how far Newton method will go. Conservative choice stays well inside $W$. $a$ between 2 and 10 moves us much closer to origin.

Remark: This rescaling is for theoretical analysis! Computationally we do not know where the origin is. In fact, the whole point of computation is to find optimal point which is origin on our picture.

We did not say how to handle equality constraints. Some possibilities are

- parametrize hyperplane $Ax + b = 0$, that is write $x = Fy + w$ with appropriate $F$ and $w$ (which can be computed using linear algebra)
- using dual problem may help

- compute Newton step via equality constrained quadratic optimization starting from feasible point. This can be done via equation solving (KKT equations are linear in this case) and leads to feasible points
- like above but starting from infeasible point and trying to get feasible one (line search may prevent this)

Given linear constraints $Ax = b$ and $x_n$ such that $Ax_n = b$ we want search direction $d_i$ such that $Ad_i = 0$. In Newton method with equality constraints we get $d_i$ solving system of equations

$$\begin{pmatrix} \nabla^2 f(x_i) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} d_i \\ w \end{pmatrix} = \begin{pmatrix} -\nabla f(x_i) \\ 0 \end{pmatrix}$$

One can show that using $d_i$ above is equivalent to replacing $f$ by $h(x) = f(Jx+s)$ where $As = b$ and $J$ is matrix such that $AJ = 0$ and $J$ has rank equal to dimension of kernel of $A$. In other words we could parametrize set of solution to $Ax = b$. Since Newton method is affine invariant convergence results for unconstrained case still work.

If initial point is infeasible, that is $Ax - b \neq 0$ we can use the following system of equations to find $d_i$ such that $x_i + d_i$ is feasible:

$$\begin{pmatrix} \nabla^2 f(x_i) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} d_i \\ w \end{pmatrix} = \begin{pmatrix} -\nabla f(x_i) \\ Ax - b \end{pmatrix}$$

## 1.5 Further reading

Stephen Boyd, Lieven Vandenberghe, Convex Optimization, chapters 5 (duality), 10, 11 (interior point method).

David G. Luenberger, Yinyu Ye, Linear and Nonlinear Programming, chapters 11, 12, 13.

A. Nemirovski, INTERIOR POINT POLYNOMIAL TIME METHODS IN CONVEX PROGRAMMING, lecture notes, chapters 3 and 4.