Convex Optimization 10-725, Lecture 14: Self-concordant function and interior point method

Yuanzhi Li

Assistant Professor, Carnegie Mellon University

Today

Last lecture

- We have learnt the Hessian Matrix, Newton's method and pre-conditioned gradient descent.

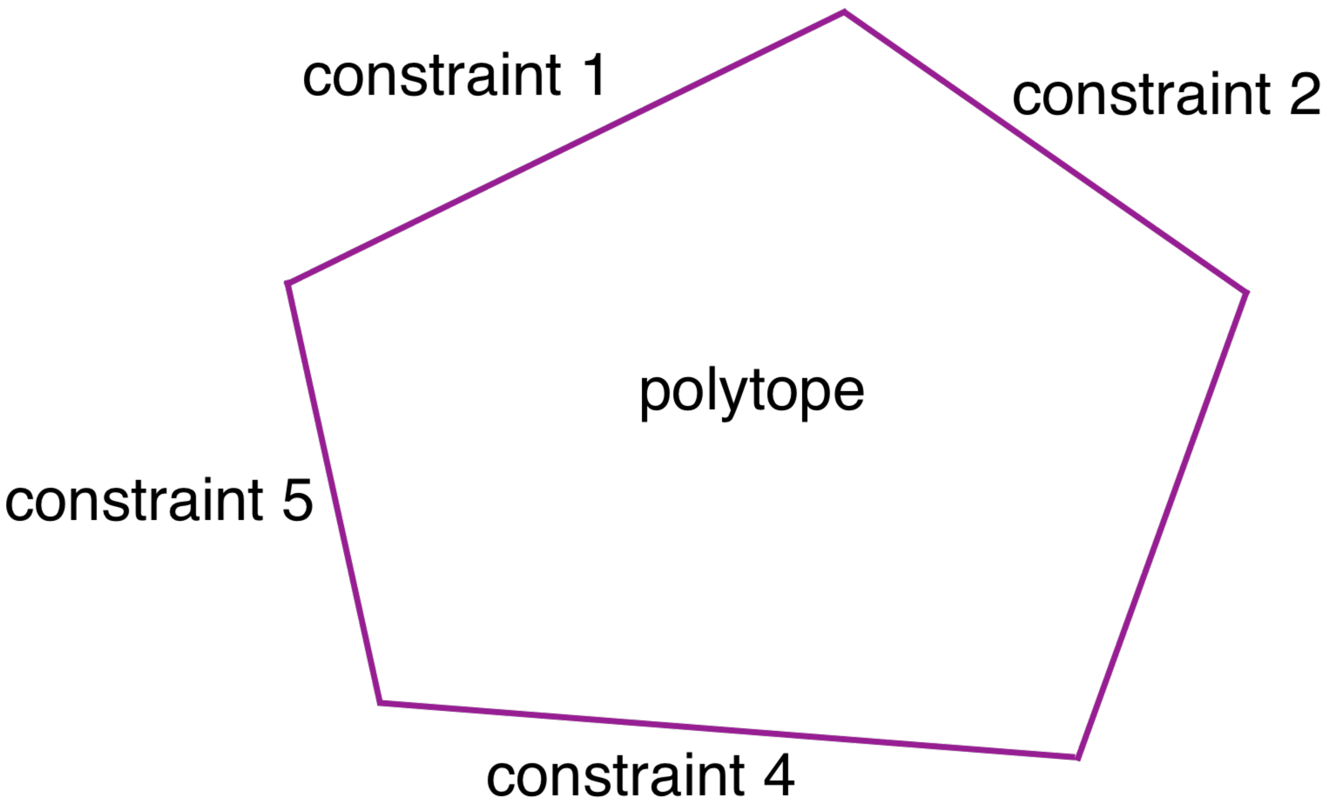
This lecture

- We are going to move further to the "geometry" side of convex optimization, and we will learn the interior point method.
- We will see an application of Newton's method / pre-conditioned gradient descent.

Motivation

- Constraint optimization: Optimizing a function $f(x)$ over a constraint set $\mathcal{D}$ which is a polytope:

$$\mathcal{D} = \{x \in \mathbb{R}^d \mid \forall i \in [n], \langle a_i, x \rangle \leq b_i \}$$



-

Motivation

- Constraint optimization: Optimizing a function $f(x)$ over a constraint set $\mathcal{D}$ which is a polytope:

$$\mathcal{D} = \{x \in \mathbb{R}^d \mid \forall i \in [n], \langle a_i, x \rangle \leq b_i \rangle\}$$

- We have learnt how to transfer it into a Minmax optimization problem:

$$\max_{\alpha_1, \cdots, \alpha_n \geq 0} \min_{x \in \mathbb{R}^d} \left(f(x) + \sum_{i \in [n]} \alpha_i (\langle a_i, x \rangle - b_i) \right)$$

- However, Minmax optimization is in general not easy, Gradient Descent Ascent might not even converge.
- Can we solve it faster and more stably?

Motivation

- This lecture we are going to see the interior point method for constraint optimization, which is theoretically very fast.
- Intuition: transfer the constraint optimization $\min_{x \in \mathcal{D}} f(x)$ for a *convex set $\mathcal{D}$* to an unconstraint optimization problem by introducing the convex barrier function, a function of type

$$R(x) = \begin{cases} = +\infty & \text{if } x \in \partial\mathcal{D}; \\ \in (-\infty, +\infty) & \text{if } x \in \mathcal{D}^o. \end{cases}$$

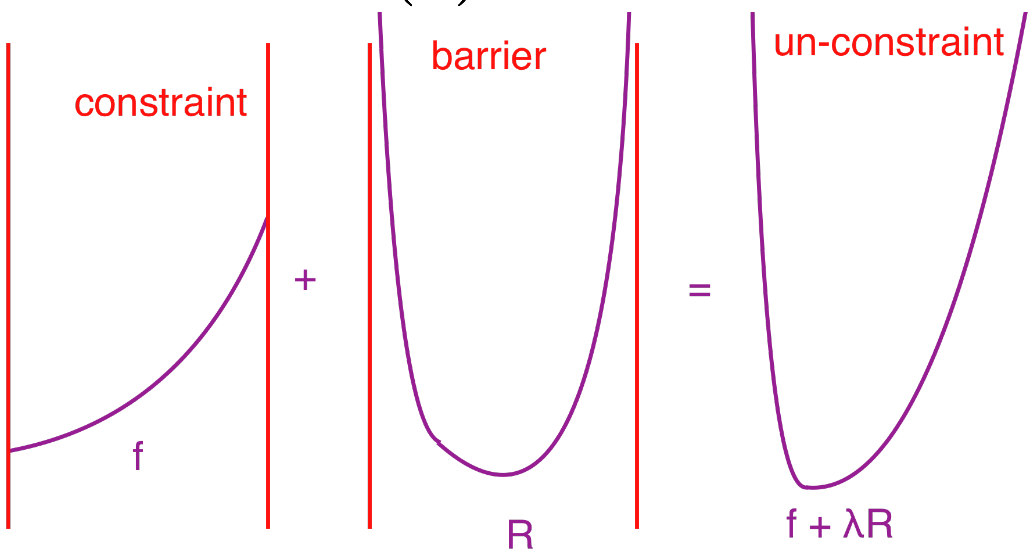
- Here, $\partial\mathcal{D}$ means the boundary of $\mathcal{D}$, $\mathcal{D}^o$ means the interior of $\mathcal{D}$.
- Minimize $f(x) + \lambda R(x)$ for a sufficiently small $\lambda > 0$ gives us *approximately* the minimizer of $f(x)$ in $\mathcal{D}$.

Motivation

- Intuition: transfer the constraint optimization $\min_{x \in \mathcal{D}} f(x)$ for a *convex set* to an unconstraint optimization problem by introducing the convex barrier function, a differentiable function of type

$$R(x) = \begin{cases} = +\infty & \text{if } x \in \partial\mathcal{D}; \\ \in (-\infty, +\infty) & \text{if } x \in \mathcal{D}^o. \end{cases}$$

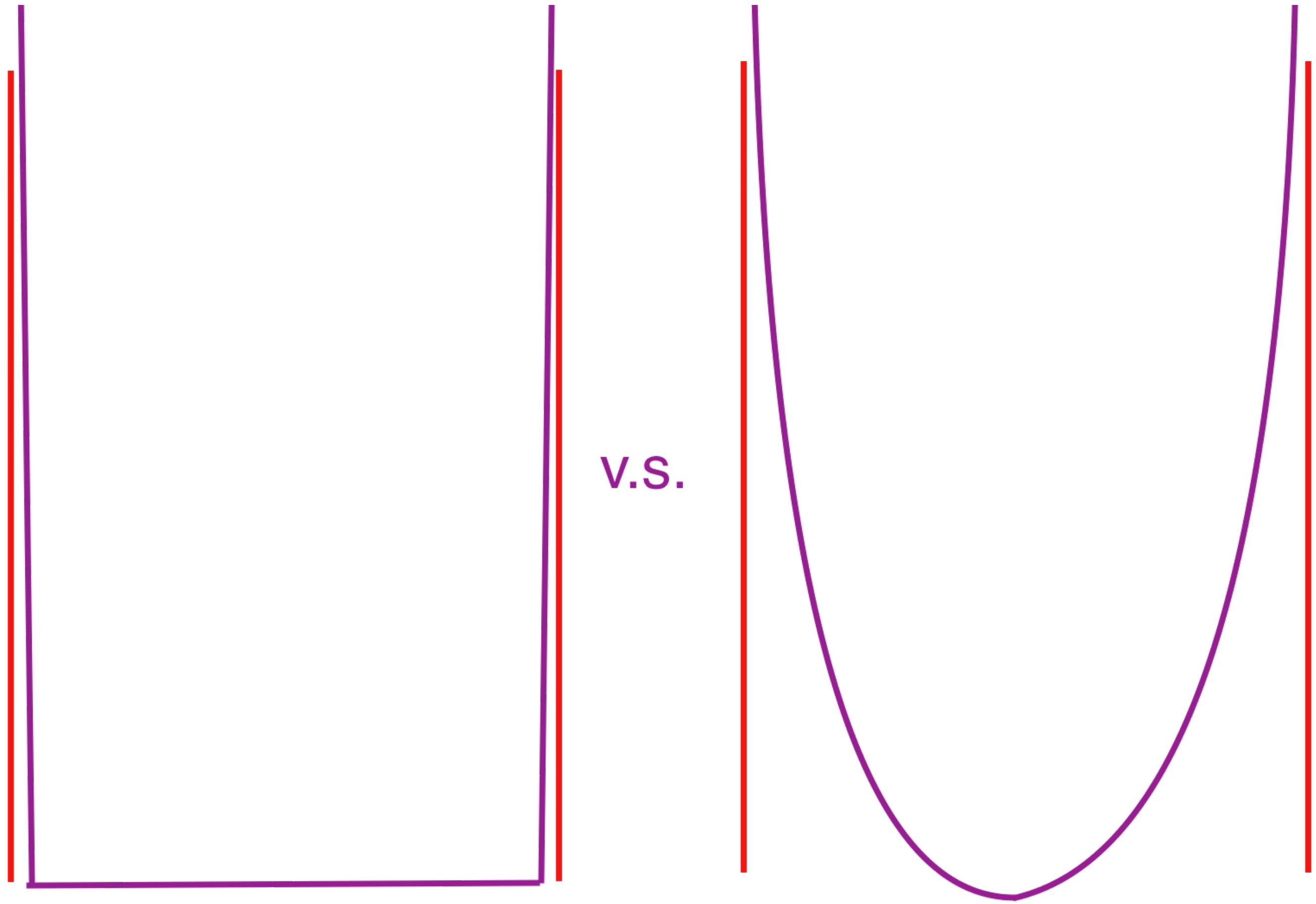
- Minimize $f(x) + \lambda R(x)$ using gradient descent, starting from a point $x \in \mathcal{D}^o$ for a sufficiently small $\lambda > 0$ gives us *approximately* the minimizer of $f(x)$ in $\mathcal{D}$.



-

Motivation

- Question: How do we design such a barrier function? What is a *good* barrier function?



-

- Which one is better?

Motivation

- How do we design such a barrier function? What is a *good* barrier function?
- Intuition: Since we are planning to run gradient descent to optimize $f(x) + \lambda R(x)$, the more smooth/Lipschitz of $R(x)$, the better.
- But $R(x) \to \infty$ when $x \to \partial \mathcal{D}$, so $R(x)$ can not be smooth/Lipschitz!
- Upper quadratic bound can not be true: for every y,

$$R(y) \leq R(x) + \langle \nabla R(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2$$

- Lipschitzness can not be true: $|R(y) - R(x)| \leq L\|x - y\|_2$.
- How do we measure the *goodness* of R?
- Let us see through some examples.

Differentiable Barrier Function

- Examples of a barrier function for

$$\mathcal{D} = \{x \in \mathbb{R}^d \mid \forall i \in [n], \langle a_i, x \rangle \leq b_i\}$$

- One such function is given by

$$R(x) = -\sum_{i \in [n]} \log(b_i - \langle a_i, x \rangle)$$

- Another function is given by

$$R(x) = \sum_{i \in [n]} \frac{1}{b_i - \langle a_i, x \rangle}$$

- Another function is given by

$$R(x) = \sum_{i \in [n]} \frac{1}{(b_i - \langle a_i, x \rangle)^2}$$

- There is no smoothness/Lipschitzness for any barrier functions, so, which one is better?

Differentiable Barrier Function: Goodness?

- Smoothness/Lipschitzness: A measure of how *consistent* the function is: how much does the function change if we change the input to this function by a little bit.
- Gradient Descent only uses local information of the function: the gradient at the current point.
- If the function remains rather *unchanged* when we change the input, then gradient descent works very well.

Differentiable Barrier Function

- Example of a barrier function for

$$\mathcal{D} = \{x \in \mathbb{R}^d \mid \forall i \in [n], \langle a_i, x\rangle \le b_i\}$$

- One such function is given by

$$R(x) = -\sum_{i\in[n]} \log(b_i - \langle a_i, x\rangle)$$

- Another function is given by

$$R(x) = \sum_{i\in[n]} \frac{1}{b_i - \langle a_i, x\rangle}$$

- Another function is given by

$$R(x) = \sum_{i\in[n]} \frac{1}{(b_i - \langle a_i, x\rangle)^2}$$

- Observation: When $\langle a_i, x\rangle \to b_i$, $\log(b_i - \langle a_i, x\rangle) \to +\infty$ much slower than the other two, so it's better because it changes slower?

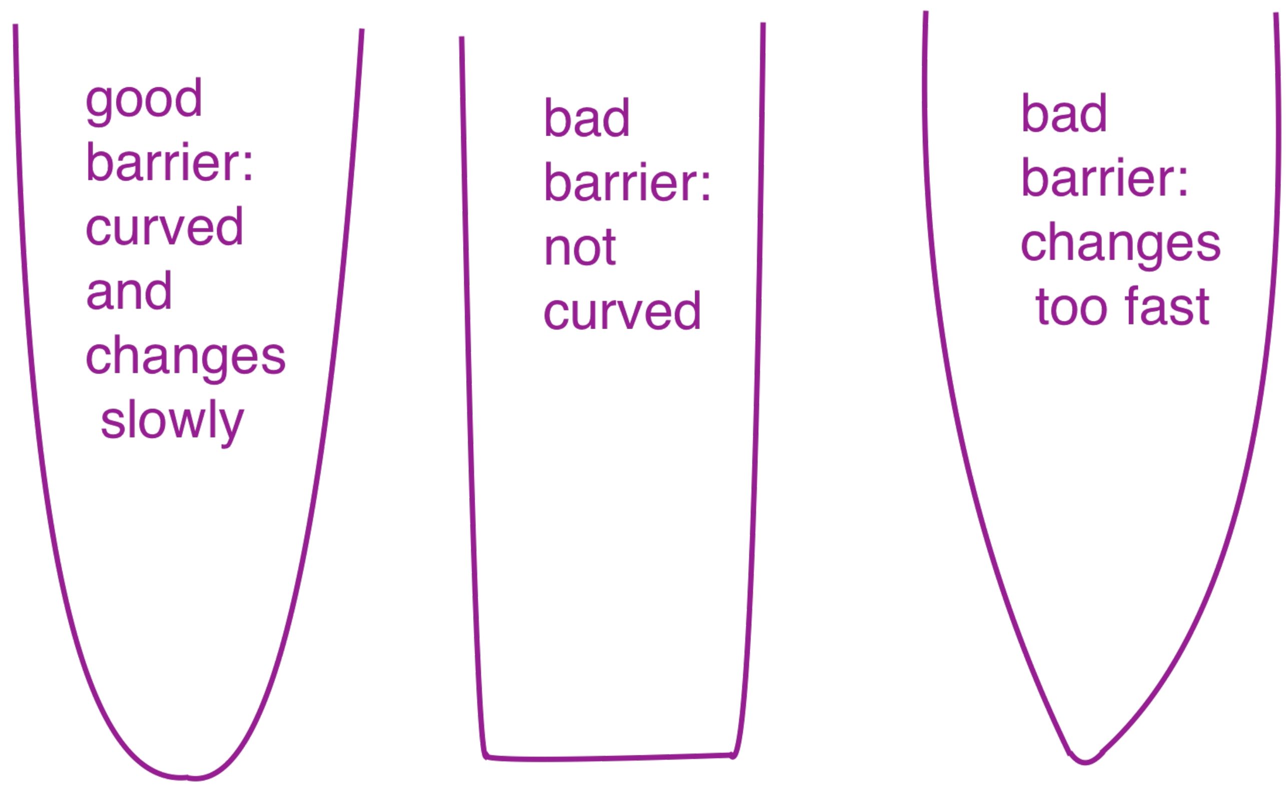
Differentiable Barrier Function

- But if we keep doing this for

$$\mathcal{D} = \{x \in \mathbb{R}^d \mid \forall i \in [n], \langle a_i, x \rangle \le b_i \rangle\}$$

- One such function is given by $R(x) = -\sum_{i \in [n]} \log(b_i - \langle a_i, x \rangle)$.
- One such function is given by $R(x) = -\sum_{i \in [n]} \log \log(b_i - \langle a_i, x \rangle)$.
- One such function is given by $R(x) = -\sum_{i \in [n]} \log \log \log(b_i - \langle a_i, x \rangle)$.
- One such function is given by $R(x) = -\sum_{i \in [n]} \log \log \log \log \log \log \log \log \log \log \log \log \log(b_i - \langle a_i, x \rangle)$.
- Eventually it becomes the indicator function which is zero for $x \in \mathcal{D}^o$ and $+\infty$ at boundary, not good.
- We need the function to have some *curve*!
- We need the function to change slowly!

Differentiable Barrier Function



-
- We need the function to have some *curve*!
- We need the function to change slowly!
- How do we depict it mathematically?

Differentiable Barrier Function: Goodness?

- Recall: Lipschitzness of a function h: $\|\nabla h(x)\|_2$ is bounded by some absolute value.
- Recall: Smoothness of a function h: $\|\nabla^2 h(x)\|_{spectral\ norm}$ is bounded absolute value.
- For a barrier function R, non of these can be bounded ...
- So we define "concordance": these values are not bounded by absolute value, but bounded by some function.
- In fact, we shall define "self-concordance": these values are bounded by the function R itself.

Self-concordant Barrier Function

- A barrier function R for a *convex* set $\mathcal{D}$ in dimension d is called self-concordant with parameter ν, if for every $x \in \mathcal{D}^o$ and for every $v \in \mathbb{R}^d$:

$$\langle \nabla R(x), v \rangle^2 \leq \nu \times v^\top \nabla^2 R(x) v$$

$$|v^\top \nabla^2 R(x) v|^{3/2} \geq \frac{1}{2} \times \nabla^3 R(x)(v, v, v)$$

- Here, $\nabla^3 R(x)(v, v, v) = \frac{d^3}{dt^3} R(x + tv) \mid_{t=0}$.
- Actually, $\langle \nabla R(x), v \rangle = \frac{d}{dt} R(x + tv) \mid_{t=0}$, $v^\top \nabla^2 R(x) v = \frac{d^2}{dt^2} R(x + tv)$.
- Why it is a $* \leq \nu *$? — We need the function to change slowly! The smaller ν is, the better.
- Why it is a $* \geq 1/2 *$? — Recall: We need the function to have some *curve*. The larger the Hessian, the more "curved" is the function.

Self-concordant Barrier Function

- A barrier function R for a *convex* set $\mathcal{D}$ in dimension d is called self-concordant with parameter $\nu \geq 0$, if for every $x \in \mathcal{D}^o$ and for every $v \in \mathcal{R}^d$:

$$\langle \nabla R(x), v \rangle^2 \leq \nu \times v^\top \nabla^2 R(x) v$$

$$|v^\top \nabla^2 R(x) v|^{3/2} \geq \frac{1}{2} \times |\nabla^3 R(x)(v, v, v)|$$

- Gradient small: change slowly, *check*.
- Hessian large: curvy. *check*.
- Seplling mistake: *chek*.
- Reason for ν and 2: these inequalities are not *scaling invariant*: $R/10$ will satisfy the first inequality more easily but the second one more difficultly.

Example of self-concordant Barrier Function

- $R(x) = -\log x$ for $x \geq 0$.
- $\frac{d}{dt} R(x+t) \mid_{t=0} = -\frac{1}{x}$.
- $\frac{d^2}{dt^2} R(x+t) \mid_{t=0} = \frac{1}{x^2}$.
- $\frac{d^3}{dt^3} R(x+t) \mid_{t=0} = -\frac{2}{x^3}$.
- Recall we need

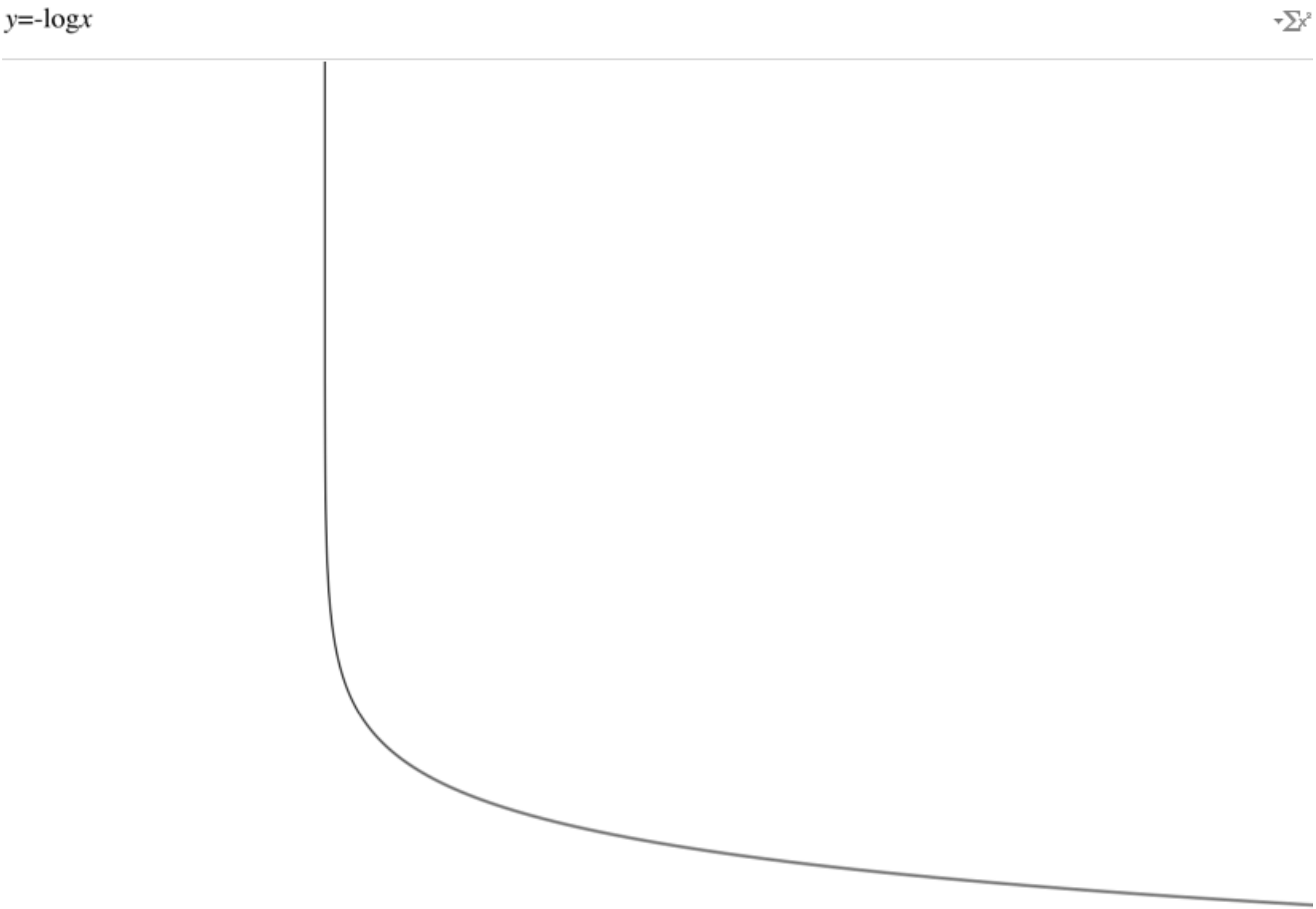
$$\langle \nabla R(x), v \rangle^2 \leq \nu \times v^\top \nabla^2 R(x) v$$

$$|v^\top \nabla^2 R(x) v|^{3/2} \geq \frac{1}{2} \times |\nabla^3 R(x)(v, v, v)|$$

- For $R(x) = -\log x$, $\nu = 1$. In fact, self-concordant barrier functions are functions exactly like $-\log x$.

Example of self-concordant Barrier Function

- In fact, self-concordant barrier functions are functions exactly like $-\log x$.



Example of self-concordant Barrier Function

- Fact: Self-concordant barrier function is convex. (Implied by $\langle \nabla R(x), v \rangle^2 \leq \nu \times v^\top \nabla^2 R(x) v$).
- Fact: If R_1, R_2 are ν self-concordant barrier function for set $\mathcal{D}$.
  - Then $R_1 + R_2$ is 2ν self-concordant barrier function for set $\mathcal{D}$.
  - Then $R_3(x) = R_1(Ax + b)$ is ν self-concordant for every matrix A and vector b.
- Fact: For every *convex* set $\mathcal{D}$ in $\mathbb{R}^d$, there is a self-concordant barrier function R for it with parameter d (the volumetric barrier function).

Examples of self-concordant Barrier Functions

- The set
$$\mathcal{D} = \{x \in \mathbb{R}^d \mid \forall i \in [n], \langle a_i, x \rangle \le b_i \rangle\}$$
- With
$$R(x) = -\sum_{i \in [n]} \log(b_i - \langle a_i, x \rangle)$$
- We have that
$$R(x + tv) = -\sum_{i \in [n]} \log(b_i - \langle a_i, x \rangle - t\langle a_i, v \rangle)$$
- We can calculate that
$$\frac{d}{dt} R(x + tv) \mid_{t=0} = \sum_{i \in [n]} \frac{\langle a_i, v \rangle}{b_i - \langle a_i, x \rangle}$$

Examples of self-concordant Barrier Functions

- We have that

$$R(x+tv) = -\sum_{i\in[n]} \log(b_i - \langle a_i, x\rangle - t\langle a_i, v\rangle)$$

- We can calculate that

$$\frac{d}{dt}R(x+tv)\mid_{t=0} = \sum_{i\in[n]} \frac{\langle a_i, v\rangle}{b_i - \langle a_i, x\rangle}$$

- We can calculate that

$$\frac{d^2}{dt^2}R(x+tv)\mid_{t=0} = \sum_{i\in[n]} \frac{\langle a_i, v\rangle^2}{(b_i - \langle a_i, x\rangle)^2}$$

- We can calculate that

$$\frac{d^3}{dt^3}R(x+tv)\mid_{t=0} = \sum_{i\in[n]} \frac{2\langle a_i, v\rangle^3}{(b_i - \langle a_i, x\rangle)^3}$$

- Thus, R is $O(n)$-self-concordant.

Interior point method

- Now we have defined a "good" barrier function R, how do we use it to minimize f over a constraint set $\mathcal{D}$?
- Why this definition of self-concordance is "good"?
- How do we pick λ in the $f(x) + \lambda R(x)$?
- We will answer these questions via the interior point method to minimize $f(x) + \lambda R(x)$.

Interior point method

- For a function f and a *convex* constraint set $\mathcal{D}$ with a ν-self-concordant barrier R,
- Interior point method: at every step:
  - Update: $\lambda_{t+1} = \lambda_t(1-\beta)$ for $\beta \in (0,1)$.
  - Find the minimizer $x^*_{t+1} = \text{argmin}\{f(x) + \lambda_{t+1}R(x)\}$ by running pre-conditioned gradient descent with learning rate η, starting from x^*_t.
- Observe: $\lambda_t \to 0$ as $t \to \infty$, so eventually we find an *approximate* minimizer of x in $\mathcal{D}$.
- Question: How do we pick β? Why pre-conditioned gradient descent works? Why self-concordant?
- We are going to see an answer in the special case of linear programming, when f is a linear function.

Interior point method

- Main theorem: When $f(x) = \langle c, x \rangle$ for a vector $c \in \mathbb{R}^d$, over a *convex* constraint set $\mathcal{D}$ with a ν-self-concordant barrier R,
- For every iteration t, define $F_{t+1}(x) = f(x) + \lambda_{t+1} R(x)$:
- For every $\beta \leq \frac{1}{16\sqrt{\nu}}$, every learning rate $\eta \leq \frac{1}{8}$
- We have that the trajectory of pre-conditioned gradient descent stays inside the Dikin's Ellipsoid:

$$\mathcal{E}_t = \left\{ x \in \mathbb{R}^d \mid (x - x_t^*)^\top \nabla^2 R(x_t^*)(x - x_t^*) \leq \frac{1}{8} \right\}$$

- Moreover, for every $x \in \mathcal{E}_t$: $F_{t+1}(x)$ is a *sandwich function*:

$$\frac{1}{4} \nabla^2 R(x_t^*) \preceq \nabla^2 F_{t+1}(x) / \lambda_{t+1} \preceq 4 \nabla^2 R(x_t^*)$$

- Starting from x_t^*, pre-conditioned gradient descent converges at a fast linear rate to x_{t+1}^*.

Interior point method

- In other words, for ν-self-concordant barrier function R, to optimize $F_{t+1}(x) = \langle c, x \rangle + \lambda_{t+1} R(x)$ for $\lambda_{t+1} \in [(1 - 1/(16\sqrt{\nu}))\lambda_t, \lambda_t]$, the trajectory of pre-conditioned gradient descent (starting from x_t^*) stays inside the *nice region*: Dikin's Ellipsoid:

$$\mathcal{E}_t = \left\{ x \in \mathbb{R}^d \mid (x - x_t^*)^\top \nabla^2 R(x_t^*)(x - x_t^*) \le \frac{1}{8} \right\}$$

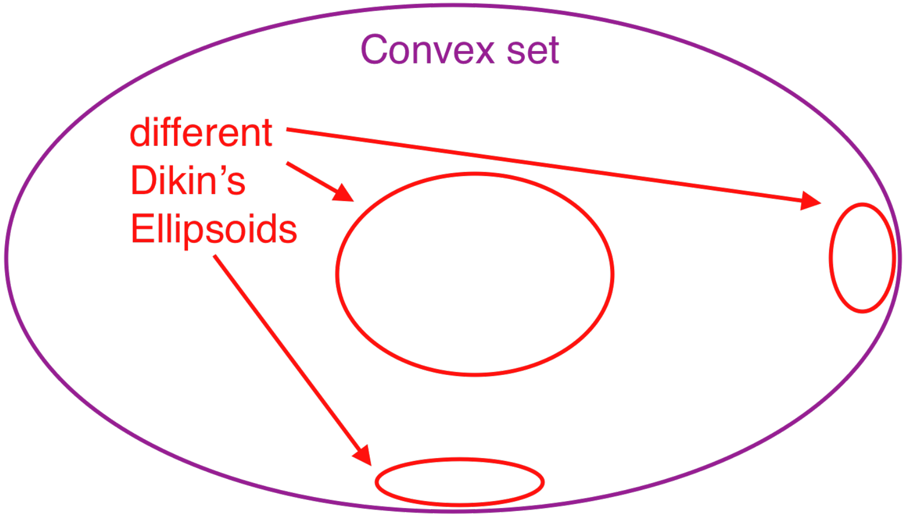
- In particular, $x_{t+1}^* \in \mathcal{E}_t$.
- In this region, for every $x \in \mathcal{E}_t$: $F_{t+1}(x)$ is a *sandwich function*:

$$\frac{1}{4}\nabla^2 R(x_t^*) \preceq \nabla^2 F_{t+1}(x)/\lambda_{t+1} \preceq 4\nabla^2 R(x_t^*)$$

Dikin Ellipsoid: Intuition

- Recall Lipschitzness of Hessian: Hessian of F stays rather unchanged if we make a small change of x in Euclidean distance.
- Self-concordance: Hessian of F stays rather unchanged when we make a small change of x in Dikin's Ellipsoid.
- Dikin's Ellipsoid is different for different point x_t^*:

$$\mathcal{E}_t = \left\{ x \in \mathbb{R}^d \mid (x - x_t^*)^\top \nabla^2 R(x_t^*)(x - x_t^*) \leq \frac{1}{8} \right\}$$

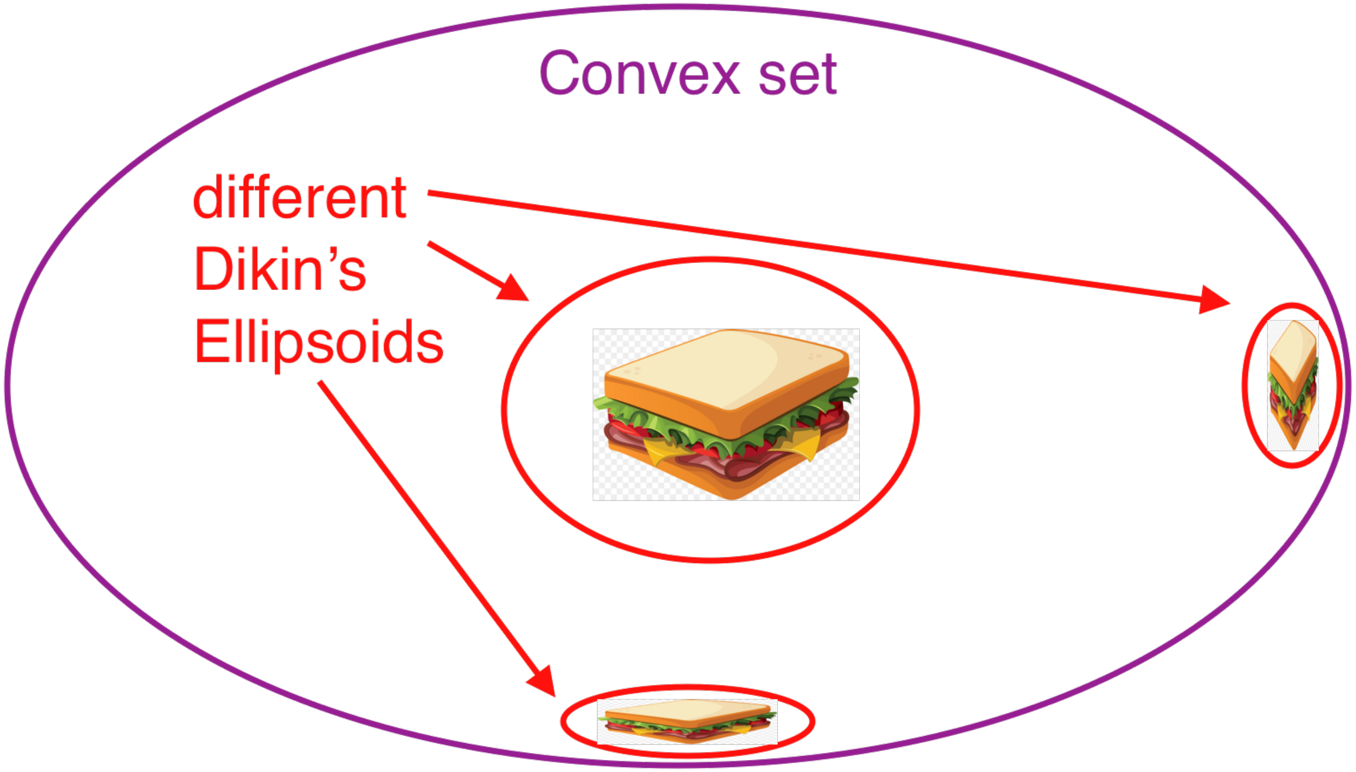


-

Dikin Ellipsoid: Intuition

- Self-concordance: Hessian of F stays rather unchanged (sandwitch) when we make a small change of x in Dikin's Ellipsoid.

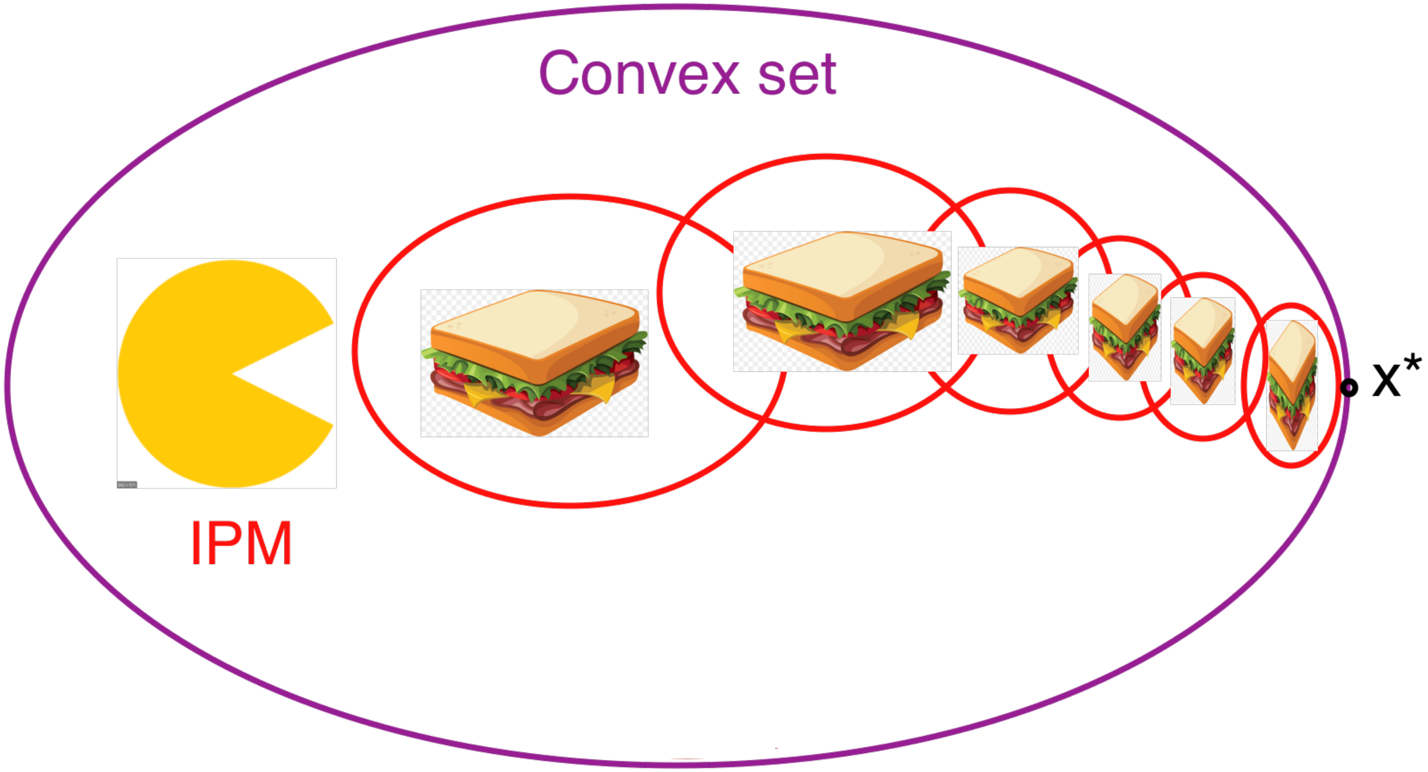
$$\frac{1}{4}\nabla^2 R(x_t^*) \preceq \nabla^2 F_{t+1}(x)/\lambda_{t+1} \preceq 4\nabla^2 R(x_t^*)$$



-

Interior Point Method: Intuition

- Sandwich eating method:



-

Interior point method: Proof

- Proof: Let me first prove that for every $z, x \in \mathcal{D}^o$, let $h = z - x$, if

$$h^\top \nabla^2 R(x) h \le \frac{1}{8}$$

- Then we have:

$$\frac{1}{4}\nabla^2 R(x) \preceq \nabla^2 R(z) \preceq 4\nabla^2 R(x)$$

- In other words: Self-concordant implies sandwich in Dikin's Ellipsoid – This is the spirit of this lecture, the very *geometry* side of convex optimization.

Interior point method: Proof

- For every vector v, let us denote $g(t) = v^\top \nabla^2 R(x + t(z - x))v$.
- Observe that

$$v^\top \nabla^2 R(x + t(z - x))v = \frac{\partial^2}{\partial r^2} R(x + t(z - x) + rv) \mid_{r=0}$$

- For $v = z - x$, we have that

$$\frac{dg}{dt} := g'(t) = \frac{\partial^3}{\partial r^3} R(x + t(z - x) + r(z - x)) \mid_{r=0}$$

$$= \nabla^3 R(x + t(z - x))(z - x, z - x, z - x)$$

- Since R is self-concordant, we have that

$$|\nabla^3 R(x+t(z-x))(z-x, z-x, z-x)| \leq \frac{1}{2}|(z-x)\nabla^2 R(x+t(z-x))(z-x)|^{3/2}$$

Interior point method: Proof

- Now we have: $g(t) = v^\top \nabla^2 R(x + t(z - x))v$.
- For $v = z - x$, we have that

$$\frac{dg}{dt} := g'(t) = \frac{\partial^3}{\partial r^3} R(x + t(z - x) + r(z - x)) \,|_{r=0}$$

$$= \nabla^3 R(x + t(z - x))(z - x, z - x, z - x)$$

- Since R is self-concordant, we have that

$$|\nabla^3 R(x+t(z-x))(z-x, z-x, z-x)| \le 2|(z-x)\nabla^2 R(x+t(z-x))(z-x)|^{3/2}$$

- This implies that

$$|g'(t)| \le 2|g(t)|^{3/2}$$

Interior point method: Proof

- Now we have: for $g(t) = v^\top \nabla^2 R(x + t(z - x))v$, $v = z - x$

$$|g'(t)| \le 2|g(t)|^{3/2}$$

- Note that $g(0) = (z - x)^\top \nabla^2 R(x)(z - x) \le \frac{1}{8}$ as our assumption.
- Key observation: $|g'(t)| \le 2|g(t)|^{3/2}, g(0) \in [0, 1/8]$ implies that for every $t \in [0, 1]$:

$$\frac{1}{2} g(0) \le g(t) \le 2g(0)$$

- Which implies that (taking $t = 1$)

$$\frac{1}{2}(z - x)^\top \nabla^2 R(x)(z - x) \le (z - x)^\top \nabla^2 R(z)(z - x)$$

$$\le 2(z - x)^\top \nabla^2 R(x)(z - x)$$

- It is a special case of what we want to show, when $v = z - x$.

Interior point method: Proof

- For every unit vector v, let us denote $g(t) = v^\top \nabla^2 R(x + t(z - x))v$.
- For general direction v, we have that:

$$g'(t) = \frac{\partial^3}{\partial r^2 \partial t} R(x + t(z - x) + rv) \mid_{r=0} = \nabla^3 R(x + t(z - x))(z - x, v, v)$$

- Fact: if for every v, w,

$$|\nabla^3 R(w)(v, v, v)| \leq 2|v^\top \nabla^2 R(w)v|^{3/2}$$

- Then

$$|\nabla^3 R(x + t(z - x))(z - x, v, v)| \leq$$

$$2|v^\top \nabla^2 R(x + t(z - x))v||(z - x)^\top \nabla^2 R(x + t(z - x))(z - x)|^{1/2}$$

Interior point method: Proof

- Let us denote $g(t) = v^\top \nabla^2 R(x + t(z - x))v$.
- Now we have

$$g'(t) = \frac{\partial^3}{\partial r^2 \partial t} R(x + t(z - x) + rv) \mid_{r=0} = \nabla^3 R(x + t(z - x))(z - x, v, v)$$

- Together with

$$|\nabla^3 R(x + t(z - x))(z - x, v, v)| \le$$

$$2|v^\top \nabla^2 R(x + t(z - x))v||(z - x)^\top \nabla^2 R(x + t(z - x))(z - x)|^{1/2}$$

- Note that we have already proved (in the special case when $v = z - x$): $|(z - x)^\top \nabla^2 R(x + t(z - x))(z - x)| \le 1/4$, this implies for every $t \in [0, 1]$

$$|g'(t)| \le |g(t)|$$

Interior point method: Proof

- Now we have for $g(t) = v^\top \nabla^2 R(x + t(z - x))v$, for every $t \in [0, 1]$

$$|g'(t)| \leq |g(t)|$$

- Since $|g(0)| \leq \frac{1}{8}$, we can also conclude that

$$\frac{1}{4} g(0) \leq g(1) \leq 4g(0)$$

- Which is

$$\frac{1}{4} v^\top \nabla^2 R(x) v \leq v^\top \nabla^2 R(z) v \leq 4 v^\top \nabla^2 R(x) v$$

- Since this is true for every v, we conclude

$$\frac{1}{4} \nabla^2 R(x) \preceq \nabla^2 R(z) \preceq 4 \nabla^2 R(x)$$

Interior point method: Proof

- We first show that $x^*_{t+1} \in \mathcal{E}_t$. To see that, notice that $x^*_{t+1} = \operatorname{argmin}\{f(x) + \lambda_{t+1} R(x)\}$ where $f = \langle c, x \rangle$, therefore we have:

$$c + \lambda_{t+1} \nabla R(x^*_{t+1}) = 0, \quad c + \lambda_t \nabla R(x^*_t) = 0$$

- Which implies that

$$\nabla R(x^*_{t+1}) = -\frac{c}{\lambda_{t+1}}, \quad \nabla R(x^*_t) = -\frac{c}{\lambda_t}$$

- Let us define $g(s) = R(x^*_t + s(x^*_{t+1} - x^*_t))$, we have that

$$g'(0) = \langle \nabla R(x^*_t), x^*_{t+1} - x^*_t \rangle, \quad g'(1) = \langle \nabla R(x^*_{t+1}), x^*_{t+1} - x^*_t \rangle$$

- This implies

$$\frac{g'(0)}{g'(1)} = \frac{\lambda_{t+1}}{\lambda_t}$$

Interior point method: Proof

- Now we have:

$$\frac{g'(0)}{g'(1)} = \frac{\lambda_{t+1}}{\lambda_t} \in \left[1 - \frac{1}{16\sqrt{\nu}}, 1\right]$$

- By self-concordance, we know that for every $s \in [0, 1]$.

$$[g'(s)]^2 \leq \nu g''(s)$$

- Let's for simplicity focus on the case when $g'(0) > 0$. Now, we have:

$$g'(1) = g'(0) + \int_0^1 g''(s) ds \geq g'(0) + \frac{1}{\nu^2} \int_0^1 g'(s)^2 ds \geq g'(0) + \frac{1}{\nu} g'(0)^2$$

- Hence

$$1 + \frac{1}{8\sqrt{\nu}} \geq \frac{g'(1)}{g'(0)} \geq 1 + \frac{1}{\nu} g'(0)$$

- This implies that $g'(0) \leq \frac{\sqrt{\nu}}{8}$, hence $g'(1) \leq g'(0) + \frac{1}{64}$, which implies that $g''(0) \leq \frac{1}{8}$ using that $g'''(s) \leq 2g''(s)^{3/2}$.