# Convex Optimization for Data Science

*Gasnikov Alexander*

gasnikov.av@mipt.ru

## Lecture 1. Foundation of Convex analysis

**October, 2016**

## Structure of a course

- Lecture 1. Foundation of Convex analysis
- Lecture 2. Convex optimization and Big Data applications
- Lectures 3. Complexity of optimization problems & Optimal methods for convex optimization problems
- Lecture 4. Stochastic optimization. Randomized methods
- Lectures 5. Primal-duality, regularization, restarts technique, mini-batch & Inexact oracle. Universal methods
- Lecture 6. Gradient-free methods. Coordinate descent

**Projects/Examples:**

https://arxiv.org/find/all/1/all:+gasnikov/0/1/0/all/0/1

## Main books:

*Polyak B.T.* Introduction to optimization. M. Nauka, 1983.
*Bertsekas D.P.* Nonlinear Programming. Belmont. MA: Athena Scientific, 1999.
*Boyd S., Vandenberghe L.* Convex optimization. – Cambridge University Press, 2004.
*Nesterov Yu.* Introduction Lectures on Convex Optimization. A Basic Course. Applied Optimization. – Springer, 2004.
*Nocedal J., Wright S.* Numerical optimization. – Springer, 2006.
*Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. – Philadelphia: SIAM, 2013.
*Bubeck S.* Convex optimization: algorithms and complexity // In Foundations and Trends in Machine Learning. – 2015. – V. 8. – no. 3-4. – P. 231–357.
*Evtushenko Yu.G.* http://www.ccas.ru/personal/evtush/p/198.pdf
https://www.youtube.com/user/PreMoLab (see courses of Yu.E. Nesterov and A.V. Gasnikov)

## Structure of Lecture 1

- Convex functions. Main properties. CVX
- Lagrange multipliers principle and vicinities
- Demyanov–Danskin's formula, sensitivity in optimization
- Dual problem. Sion–Kakutani theorem (min max = max min)
- KKT-theorem. The role of convexity (sufficient condition)

**Examples:**

- Conic duality – dual representation (Robust optimization)
- Constrained primal problem leads to unconstrained dual one
- Implicit primal problem leads to explicit dual one
- Non separable large-dimensional primal problem leads to small dimensional dual problem on a bounded convex set (Slater's arguments)

- **Convex set**

$$\forall\, \alpha \in [0,1], x, y \in Q \to \alpha x + (1-\alpha) y \in Q$$

**Convex function**

$$\forall\, \alpha \in [0,1], x, y \to f\left(\alpha x + (1-\alpha) y\right) \le \alpha f(x) + (1-\alpha) f(y)$$

**Examples of convex functions**

**i)** $f(x) = \ln\left(\sum_{k=1}^{n} \exp\left(\langle a_k, x\rangle\right)\right)$ // by direct investigation of the Hessian

**ii)** $f(x) = x_{[1]} + \ldots + x_{[k]}$ (sum of largest $k$ entries)

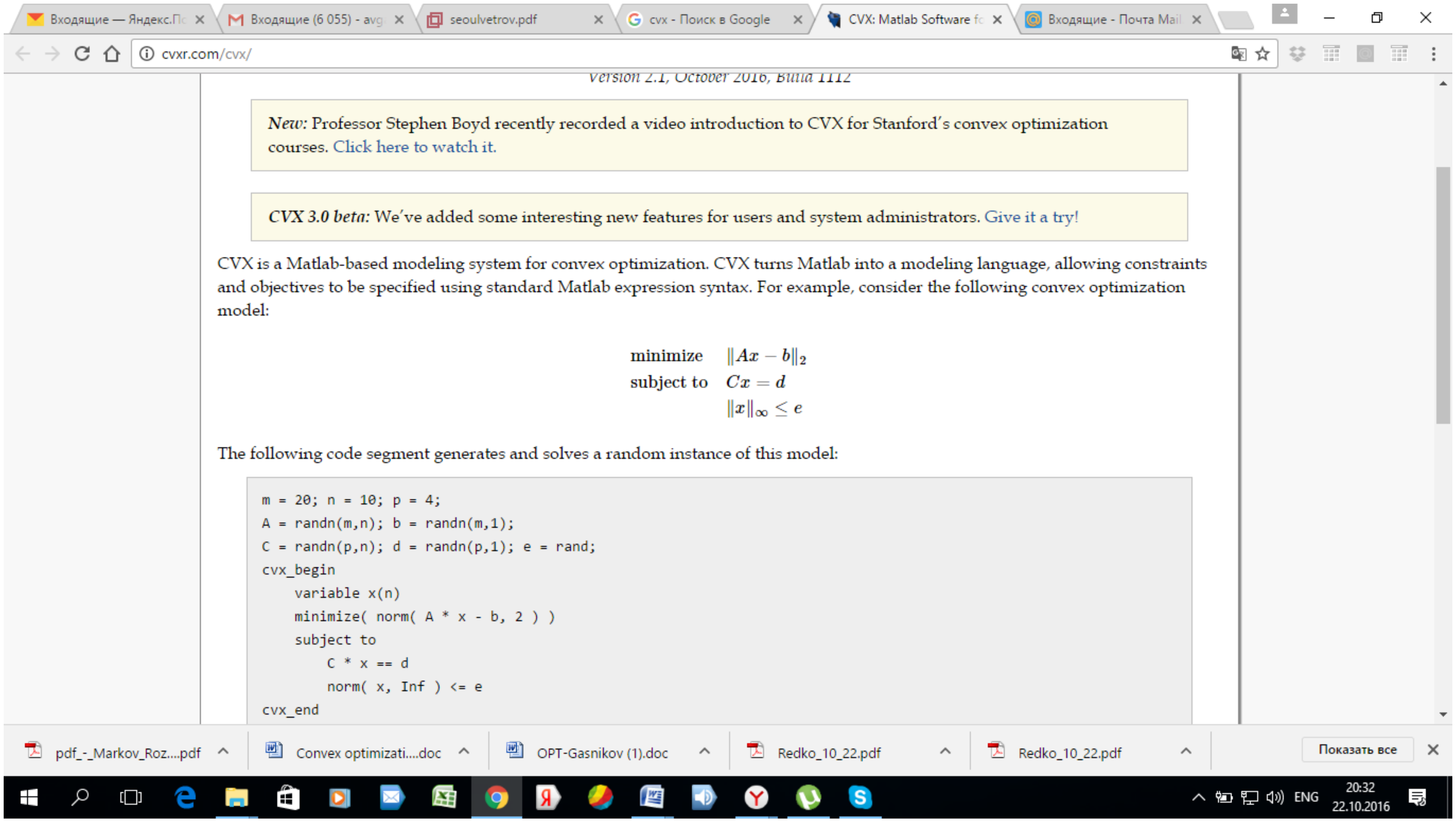
**iii)** $f(X) = \lambda_{\max}(X),\ X = X^T$

**iv)** $f(X) = \ln\det X^{-1},\ X \succ 0$

**v)** $f(x,Y) = \langle x, Y^{-1}x\rangle,\ Y \succ 0$

# Main Toolbox:

## [http://cvxr.com/cvx/](http://cvxr.com/cvx/)



*New:* Professor Stephen Boyd recently recorded a video introduction to CVX for Stanford's convex optimization courses. Click here to watch it.

*CVX 3.0 beta:* We've added some interesting new features for users and system administrators. Give it a try!

CVX is a Matlab-based modeling system for convex optimization. CVX turns Matlab into a modeling language, allowing constraints and objectives to be specified using standard Matlab expression syntax. For example, consider the following convex optimization model:

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|_2 \\ \text{subject to} & Cx = d \\ & \|x\|_\infty \le e \end{array}$$

The following code segment generates and solves a random instance of this model:

```
m = 20; n = 10; p = 4;
A = randn(m,n); b = randn(m,1);
C = randn(p,n); d = randn(p,1); e = rand;
cvx_begin
    variable x(n)
    minimize( norm( A * x - b, 2 ) )
    subject to
        C * x == d
        norm( x, Inf ) <= e
cvx_end
```

- CVX can solve the following convex optimization problems

$$\langle c, x\rangle + \langle d, y\rangle \to \min_{(x,y):\, A\begin{bmatrix} x \\ y \end{bmatrix} = b,\, \begin{bmatrix} x \\ y \end{bmatrix} \in K},$$

where $K$ – is a product of convex cones. Typically these cones: $\mathbb{R}^n_+$, $S^n_+$, $L^n_2$ (positive cone, positive semidefinite cone, Lorentz cone).

- Many convex functions have cone representation (Nesterov–Nemirovski)

$$\boxed{f(x) = \min_{y:\, A\begin{bmatrix} x \\ y \end{bmatrix} = b,\, \begin{bmatrix} x \\ y \end{bmatrix} \in K} \langle c, x\rangle + \langle d, y\rangle}\text{! (e.g. i) – v))}$$

**Other Toolboxes:**

- https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/ CPLEX (IBM product): Large-scale LP, QP
- https://www.mosek.com/ Large-scale convex optimization problems

## How to obtain convex functions?

**Lemma 1.** *Let* $f(x)$, $g(x)$ *convex,* $h(z)$ *convex increasing. Then* $\alpha f(x)+\beta g(x)$ $(\alpha,\beta\geq 0)$, $f(Ay+b)$, $h\left(f(x)\right)$ *are also convex functions.*

**Lemma 2.** *Let* $G(x,y)$ *– is convex function as a function of* $x$ *for all* $y\in Y$. *Assume that problem* $\max\limits_{y\in Y} G(x,y)$ *is solvable for all* $x$. *Then* $f(x)=\max\limits_{y\in Y} G(x,y)$ *is convex.* **Example.** $f(x)=\|Ax-b\|_{\infty}$ is convex.

**Lemma 3.** *Let* $G(x,y)$ *– is convex function as a function of* $x$ *and* $y$ *on the convex set* $Q$. *Assume that problem* $\min\limits_{y:(x,y)\in Q} G(x,y)$ *is solvable for all* $x$. *Then* $f(x)=\min\limits_{y:(x,y)\in Q} G(x,y)$ *is convex.* **Example.** $f(x)=\inf_{y\in Q}\|x-y\|$ (where $Q$ is a convex set) is convex. How can be calculated $\partial f(x)$?

## Demyanov–Danskin's formula

Let $G(x,y)$ and

$$\boxed{f(x)=\max_y G(x,y)}\Big(f(x)=\min_y G(x,y)\Big)$$

are smooth enough functions. Assume that there exists $y(x)$ such that

$$G(x,y(x))=\max_y G(x,y)\Big(G(x,y(x))=\min_y G(x,y)\Big).$$

Then $\boxed{\nabla f(x)=\nabla_x G(x,y(x))}=\left\{\dfrac{\partial G(x,y)}{\partial x_i}\right\}_i\Bigg|_{y=y(x)}$. // $f_x=G_x+\underbrace{G_y}_{0}y_x=G_x$

$$\boxed{\partial f(x)=\text{convex hull}\bigcup_{\tilde{x}:\,y(\tilde{x})=y(x)}\partial_x G(\tilde{x},y(\tilde{x}))}\ \text{(convex case).}$$

## Schur's complement

Using Lemma 3 and Demyanov–Danskin's formula

$$G(x,y) = \langle x, Cx \rangle + \langle y, Ay \rangle + 2\langle Bx, y \rangle,$$

one can show that if $A \succ 0$ (strictly, i.e. $\forall\ x \neq 0 \to \langle x, Ax \rangle > 0$) and

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succ 0. \text{ // this means that } G(x,y) \text{ is convex}$$

Then

$$C - B^T A^{-1} B \succ 0.$$

Indeed here we can find explicitly $y(x) = \arg\min_y G(x,y) = -A^{-1}Bx$. Hence

$$f(x) = \min_y G(x,y) = G(x, y(x)) = \langle x, (C - B^T A^{-1} B) x \rangle$$

is a convex function due to Lemma 3.

## Lagrange multipliers principle and Implicit function theorem

We have a sufficiently smooth optimization problem

$$f(x,y) \to \min_{g(x,y)=0},$$

where $g:\mathbb{R}^{n+m} \to \mathbb{R}^m$. Assume that implicit function theorem can be applied for $g(x,y)=0$. That is, there exists smooth $y(x)$ such that $g(x,y(x)) \equiv 0$. If $(x_*,y_*)$ is a solution of initial optimization problem then

$$\exists\, \lambda : \nabla_x L(x_*,y_*,\lambda)=0, \nabla_y L(x_*,y_*,\lambda)=0, \qquad (*)$$

where

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

$\lambda$ can be found from $g(x_*(\lambda), y_*(\lambda)) = 0$ where $x_*(\lambda), y_*(\lambda)$ satisfy (*).

## Implicit function theorem

$$g\left(x, y(x)\right) \equiv 0 \Rightarrow g_x + g_y y_x \equiv 0 \Rightarrow y_x = -g_y^{-1} g_x$$

## Fermat principle

$$\frac{d}{dx} f\left(x, y(x)\right) = 0 \Rightarrow f_x + f_y y_x = 0 \Rightarrow f_x - f_y g_y^{-1} g_x = 0$$

## Lagrange multipliers principle

$$\left.\begin{array}{l} 0 = L_x = f_x + \lambda^T g_x \\ 0 = L_y = f_y + \lambda^T g_y \;\Big\}\times\left(-g_y^{-1} g_x\right) \end{array}\right\} + \Rightarrow$$

$$f_x - f_y g_y^{-1} g_x = 0. \qquad (**)$$

Equality (**) with $g\left(x, y\right) = 0$ allows us to find extremum (optimal point).

# Sensitivity in optimization

We have a sufficiently smooth optimization problem

$$f(x) \to \min_{g(x)=b},$$

where $g : \mathbb{R}^n \to \mathbb{R}^m$ ($m < n$). Let

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

For optimal solution $x(b)$ there exists such $\lambda(b)$ that

$$L_x\big(x(b),\lambda(b)\big) = 0. \ // \ g\big(x(b)\big) = b \qquad (*)$$

$$\boxed{\nabla F(b) = \lambda(b)}, \ \boxed{F(b) = \min_{g(x)=b} f(x)}.$$

Indeed, $F(b) = L\big(x(b),\lambda(b),b\big)$; $F_b = L_x x_b + L_\lambda \lambda_b + L_b = L_b = \lambda(b)$ due to (*).

## Lagrange multipliers principle and Separation theorem (Convex case)

We have a convex optimization problem ($g:\mathbb{R}^n \to \mathbb{R}^m$)

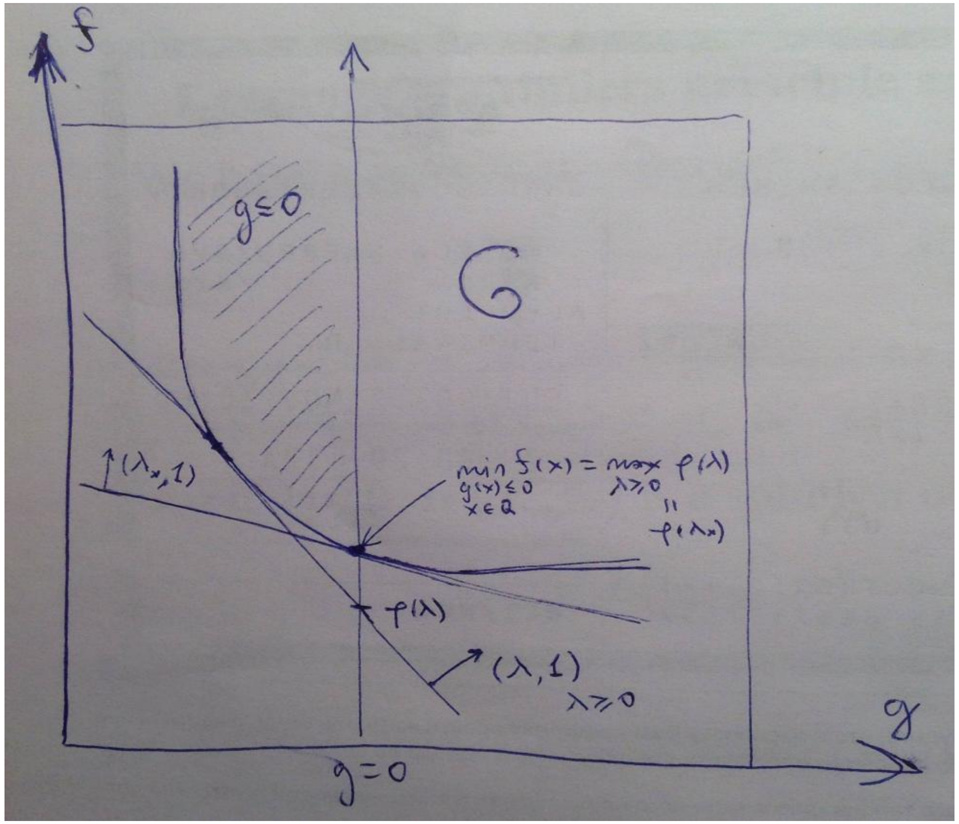
$$f(x) \to \min_{g(x)\le 0,\ x\in Q}.$$

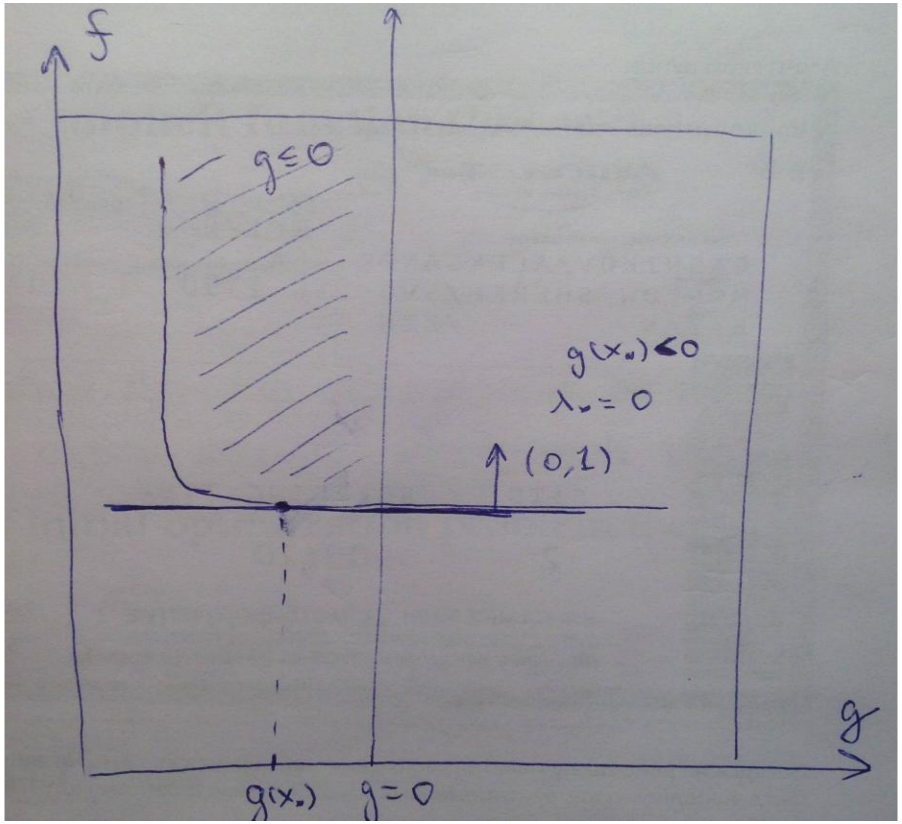
The (Pareto) set $G = \left\{\left(g(x), f(x)\right),\ x\in Q\right\} \oplus \mathbb{R}_+^{m+1}$ is close convex set in the space of pairs $(g, f)$ (http://stanford.edu/~boyd/cvxbook/bv_cvxslides.pdf).

i) $f + g\lambda = \varphi(\lambda)$, where $\lambda \ge 0$ – be a tangent hyperplane to $G$ iff

$$\varphi(\lambda) = \min_{x\in Q}\left\{ f(x) + \left\langle \lambda, g(x)\right\rangle\right\} \text{ (dual function).}$$

ii) Lagrange multipliers principle can be obtained from the Separation theorem for the set $G$ and hyperplane from i).




iii) For arbitrary $x \in Q,\ g(x) \leq 0,\ \lambda \geq 0$

$$\boxed{f(x) \geq \varphi(\lambda)}$$

$$\underbrace{\min_{g(x)\le 0,\, x\in Q} f(x)}_{\substack{primal\\ problem}} \ge \underbrace{\max_{\lambda\ge 0}\varphi(\lambda)}_{\substack{dual\\ problem}}. \text{ (weak duality)}$$

Typically for convex problem we have

$$\boxed{\min_{g(x)\le 0,\, x\in Q} f(x) = \max_{\lambda\ge 0}\varphi(\lambda)}. \text{ (strong duality)}$$

**Example (LP).** $\langle c, x\rangle \to \min\limits_{Ax=b,\, x\ge 0}$. Dual problem: $\langle b,\lambda\rangle \to \max\limits_{c-A^T\lambda\ge 0}$.

Hint: $\min\limits_{x\ge 0}\left\{\langle c,x\rangle + \max\limits_{\lambda}\langle b-Ax,\lambda\rangle\right\} = \max\limits_{\lambda}\left\{\langle b,\lambda\rangle + \min\limits_{x\ge 0}\langle c-A^T\lambda, x\rangle\right\}$.

If primal problem is compatible then we have strong duality (otherwise dual problem reach infinity).

What are the sufficient conditions for there is no duality gap (strong duality)?

**Strong duality (for convex problem):** there exists non-vertical supporting hyperplane at $(g, f) = (0, f_*)$.

## Slater’s condition

We introduce

$$Q_{\bar{\lambda}} = \left\{ \lambda \in \mathbb{R}_+^m : \varphi(\lambda) \ge \varphi(\bar{\lambda}) \right\}.$$

Assume that Slater’s condition is true (sufficient condition in KKT):

*there exists such* $\bar{x} \in Q$ *that* $g(\bar{x}) < 0$ *(*$\gamma = \min\limits_{i=1,\dots,m} \left\{ -g_i(\bar{x}) \right\}$*).*

Then

$$\boxed{\|\lambda_*\|_1 \le \max_{\lambda \in Q_{\bar{\lambda}}} \|\lambda\|_1 = \max_{\lambda \in Q_{\bar{\lambda}}} \sum_{i=1}^m |\lambda_i| \le \frac{1}{\gamma}\left( f(\bar{x}) - \varphi(\bar{\lambda}) \right)}.$$

**Hint:** $\varphi(\bar{\lambda}) \le \varphi(\lambda) = \min\limits_{x \in Q} \left\{ f(x) + \sum\limits_{i=1}^m \lambda_i g_i(x) \right\} \le f(\bar{x}) + \sum\limits_{i=1}^m \lambda_i g_i(\bar{x}).$

## Quadratic Programming

Let’s consider quadratic convex ($C \succ 0$) optimization problem ($Ax \le b$ is solvable)

$$\langle x, Cx \rangle + \langle d, x \rangle \to \min_{Ax \le b}.$$

Dual problem (we have strong duality because of affine restrictions)

$$\langle \lambda, \tilde{C}\lambda \rangle + \langle \tilde{d}, \lambda \rangle + \tilde{c} \to \min_{\lambda \ge 0},$$

$$\tilde{C} = AC^{-1}A^T,\ \tilde{d} = AC^{-1}d + b,\ \tilde{c} = \frac{1}{2}\langle d, C^{-1}d \rangle.$$

The solutions of primal and dual problems ($x_*$ and $\lambda_*$) satisfy

$$Cx_* + A^T\lambda_* + d = 0.$$

Typically, one can explicitly build dual problem iff one can explicitly connect primal and dual variables.

## Dual relaxation (strong duality)

For the problem ($A \nsucc 0,\ A^T = A$)

$$f(x) = \langle x, Ax \rangle \to \min_{\|x\|_2^2 \le 1}$$

we have the following dual problem ($C \succ 0$ means $\forall\ x \to \langle x, Cx \rangle \ge 0$)

$$\varphi(\lambda) = \begin{cases} -\lambda, \ A + \lambda I \succ 0 \\ -\infty, \ \text{otherwise} \end{cases} \to \max_{\lambda}.$$

There is no duality gap in this situation. That is we have strong duality

$$\min_{x \in Q} f(x) = \max_{\lambda} \varphi(\lambda)$$

**Note:** $f_* = \lambda_{\min},\ x_*$ – eigen vector that corresponds $\lambda_{\min}$.

## Dual relaxation (weak duality)

Consider NP-hard two-way partitioning problem

$$f\left(x\right)=\left\langle x,Wx\right\rangle \to \min_{x_i^2=1,\ i=1,\dots,n}.$$

The dual problem is

$$\varphi\left(\lambda\right)=\begin{cases}-\sum_{i=1}^{n}\lambda_i,\ W+Diag\left\{\lambda_i\right\}_{i=1}^{n}\succ 0\\ -\infty,\ \text{otherwise}\end{cases}\to \max_{\lambda}.$$

Using weak duality one can show that

$$\min_{x_i^2=1,\ i=1,\dots,n}\left\langle x,Wx\right\rangle \ge n\lambda_{\min}\left(W\right).$$

## Necessary and sufficient conditions for convex programming

We have a convex optimization problem

$$f\left(x\right) \to \min_{g\left(x\right)\le 0,\ Ax=b,\ x\in Q}.$$

Introduce: $$L\left(x,\lambda_0,\lambda,\mu\right) = \lambda_0 f\left(x\right) + \left\langle \lambda, g\left(x\right)\right\rangle + \left\langle \mu, Ax-b\right\rangle.$$

Initial problem can be equivalently reformulated as

$$\sup_{\lambda\ge 0,\ \mu} L\left(x,1,\lambda,\mu\right) \to \min_{x\in Q}.$$

Assume that: $$\min_{x\in Q} \sup_{\lambda\ge 0,\ \mu} L\left(x,1,\lambda,\mu\right) = \sup_{\lambda\ge 0,\ \mu} \min_{x\in Q} L\left(x,1,\lambda,\mu\right).$$

Then $$\min_{x\in Q} L\left(x,1,\lambda,\mu\right) \to \sup_{\lambda\ge 0,\ \mu}.$$ (dual problem)

## Complementary slackness

$$L\left(x,\lambda_0,\lambda,\mu\right)=\lambda_0 f\left(x\right)+\left\langle\lambda,g\left(x\right)\right\rangle+\left\langle\mu,Ax-b\right\rangle,$$

$$\varphi\left(\lambda,\mu\right)=\min_{x\in Q}L\left(x,1,\lambda,\mu\right)=L\left(x\left(\lambda,\mu\right),1,\lambda,\mu\right)\to\sup_{\lambda\geq 0,\,\mu}.$$

From the optimality conditions: $\varphi_\lambda=0$ or $\lambda_*=0$, $\varphi_\lambda\leq 0$; $\varphi_\mu=0$. Since

$$\varphi_\lambda=L_x x_\lambda+L_\lambda=L_\lambda=g;\ \varphi_\mu=L_x x_\mu+L_\mu=L_\mu=Ax\left(\lambda,\mu\right)-b$$

one can obtain that

$$A\cdot x\left(\lambda,\mu\right)=b,\ g\left(x\left(\lambda,\mu\right)\right)\leq 0.$$

Solution of the dual problem $\lambda_*>0$ only if $g\left(x\left(\lambda_*,\mu_*\right)\right)=0$.

Let's remove $Ax=b$, $\lambda_0=0$ means that supporting hyperplane to $G$ at $\left(g,f\right)=\left(0,f_*\right)$ is vertical.

## Karush–Kuhn–Tucker theorem (KKT)

$$f\left(x\right)\to \min_{g\left(x\right)\le 0,\ Ax=b,\ x\in Q} \ (g:\mathbb{R}^n\to\mathbb{R}^m)$$

**Necessary condition.** *Let* $x_* \in Q$ *– is a solution of the problem.*

*Then there exist* $\boxed{\lambda_0 \ge 0,\ \lambda_i \ge 0,\ i=1,\dots,m}$, $\mu$ *such that:*

*i)* $Ax_* = b$, $\boxed{\lambda_i g_i\left(x_*\right)=0}$, $i=1,\dots,m$; *(complementary slackness)*

*ii)* $\boxed{\min_{x\in Q} L\left(x,\lambda_0,\lambda,\mu\right)=L\left(x_*,\lambda_0,\lambda,\mu\right)}$.

**Sufficient condition.** *Let* $x_*$ *satisfy* $g\left(x_*\right)\le 0$, $Ax_*=b$, $x_*\in Q$ *and i), ii) with* $\lambda_0>0$ *(equivalent* $\lambda_0=1$*) then* $x_*$ *– is a solution of the problem.*

## Saddle point ($\min\max = \max\min$)

We say that $(x_*, \lambda_*)$ is a saddle point of $L(x,\lambda)$ iff for all admissible $(x,\lambda)$

$$L(x,\lambda_*) \geq L(x_*,\lambda_*) \geq L(x_*,\lambda).$$

$$\overline{L}(x) = \sup_{\lambda\in\Lambda} L(x,\lambda) \to \min_{x\in X} \text{ (i)} \qquad \underline{L}(\lambda) = \inf_{x\in X} L(x,\lambda) \to \max_{\lambda\in\Lambda} \text{ (ii)}$$

1) $\max_{\lambda\in\Lambda} \underline{L}(\lambda) \leq \min_{x\in X} \overline{L}(x)$ (if there exists saddle point then we have "="); 

2) Saddle point set of $L(x,\lambda)$ coincide with pairs of solutions (i), (ii).

## Sion–Kakutani minimax theorem

*Let continuous function $L(x,\lambda)$ is convex in $x$ and concave in $\lambda$. The sets $X$ and $\Lambda$ are convex and $X$ is compact. Then* $\max_{\lambda\in\Lambda} \underline{L}(\lambda) = \min_{x\in X} \overline{L}(x)$.

## Neyman–Pearson's Lemma

Assume we have random samples $x=\left(x_1,...,x_n\right)$ and two hypotheses about the probability nature of these samples ($L\left(\cdot\right)$ is likelihood function)

$$H_0:\ L\left(x\middle|H_0\right);\ H_1:\ L\left(x\middle|H_1\right).$$

Let's introduce decision rule: $0\le\varphi\left(x\right)\le1$ – probability to decide in favor of hypothesis $H_1$ if one observe vector of samples $x$. We'd like to find the best $\varphi\left(x\right)$ in terms of the following infinite dimensional LP-problem:

$$\beta=P\left(H_0\middle|H_1\right)=1-\int\varphi\left(x\right)L\left(x\middle|H_1\right)dx\to\min_{P\left(H_1\middle|H_0\right)\le\alpha};$$

$$P\left(H_1\middle|H_0\right)=\int\varphi\left(x\right)L\left(x\middle|H_0\right)dx\ \left|\lambda>0,\ 0\le\varphi\left(x\right)\le1.\right.$$

We can solve this problem by means of Lagrange multipliers principle

$$L\left(\varphi(\cdot),\lambda\right)=1-\int\varphi(x)L\left(x\middle|H_1\right)dx+\lambda\cdot\left(\int\varphi(x)L\left(x\middle|H_0\right)dx-\alpha\right)\to\min_{0\le\varphi(x)\le 1},$$

$$\int\varphi(x)\left(\lambda L\left(x\middle|H_0\right)-L\left(x\middle|H_1\right)\right)dx\to\min_{0\le\varphi(x)\le 1},$$

$$\varphi(x)=\begin{cases}1, & \Lambda(x)>\lambda\\ p(x), & \Lambda(x)=\lambda,\\ 0, & \Lambda(x)<\lambda\end{cases}\quad \Lambda(x)=\frac{L\left(x\middle|H_1\right)}{L\left(x\middle|H_0\right)},$$

where $$\int\limits_{\Lambda(x)>\lambda}L\left(x\middle|H_0\right)dx+\int\limits_{\Lambda(x)=\lambda}p(x)L\left(x\middle|H_0\right)dx=\alpha.$$

Multiplier $\lambda$ is determines from here in unique manner, and $\beta$ doesn't depend on the choice of $p(x)$. For more general facts about hypothesis testing via optimization see http://www2.isye.gatech.edu/~nemirovs/StatOpt_LN.pdf

## Robust Optimization (A. Nemirovski)

$$\langle c, x\rangle \to \min_{\langle \alpha, x\rangle \le \beta,\ \alpha \in \mathrm{A}} .$$

The set of vectors $\mathrm{A}$ can be infinitely big (e.g. in robust optimization $\mathrm{A}$ is a box). So in general we have a LP problem with infinitely many constraints. Can we solve this problem efficiently? The answer is YES if the set $\mathrm{A}$ has a Fourier–Motzkin representation. This mean the following

$$\mathrm{A} = \left\{\alpha:\ \exists\ u:\ A\alpha + Bu + b = 0;\ C\alpha + Du + e \in K\right\},$$

where $K$ is a convex cone (with simple enough dual cone $K^*$).

Note that all reasonable convex set have such representation, e.g. boxes,

$$X = \left\{x \in \mathbb{R}^3_+ :\ x_1 x_2 x_3 x_4 \ge 1\right\},\ X = \left\{X \in \mathbb{R}^{m\times n} :\ \|X\|_{\text{nuclear}} \le 1\right\}.$$

# Farkas' Lemma (& Conic duality)

*Assume* $\exists\ (\bar{\alpha},\bar{u})$: $A\bar{\alpha}+B\bar{u}+b=0$; $C\bar{\alpha}+D\bar{u}+e\in \operatorname{icr} K$ *or* $K=\mathbb{R}^m_+$. *Then*

$$\langle \alpha, x\rangle \le \beta \ \textit{ for all } \alpha\in \mathrm{A} \ \textbf{\textit{iff}}$$

$$A^T\mu + C^T\lambda + x = 0,\ B^T\mu + D^T\lambda = 0,\ \langle b,\mu\rangle + \langle e,\lambda\rangle \le \beta$$

*is compatible in* $(\mu,\lambda)$, *where* $\lambda\in K^* = \{\lambda:\ \langle \lambda, y\rangle \ge 0 \text{ for all } y\in K\}$.

**Show the transition "$\Leftarrow$".** For all $\alpha\in\mathrm{A}$, $\lambda\in K^*$ we have

$$0=\langle A\alpha+Bu+b,\mu\rangle,\ 0\le\langle C\alpha+Du+e,\lambda\rangle.$$

Hence

$$\langle \underbrace{-A^T\mu - C^T\lambda}_{x},\alpha \rangle + \langle \underbrace{-B^T\mu - D^T\lambda}_{0}, u\rangle \le \underbrace{\langle b,\mu\rangle + \langle e,\lambda\rangle}_{\le\beta}.$$

Using Farkas' lemma we can reformulate initial LP optimization problem:

$$\langle c, x \rangle \to \min_{(x,\mu,\lambda)\in X},$$

where the convex set $X$ has the following Fourier–Motzkin representation (reduced to cone representation, since that the problem can be efficiently solved for example by S. Boyd's CVX)

$$X = \{(x, \mu, \lambda):\ A^T \mu + C^T \lambda + x = 0,$$

$$B^T \mu + D^T \lambda = 0,\ \langle b, \mu \rangle + \langle e, \lambda \rangle \le \beta,\ \ \lambda \in K^*\}.$$

Details can be found in the book:

*Ben-Tal A., Ghaoui L.El., Nemirovski A.* Robust optimization. – Princeton University Press, 2009.

## Arbitrage free theorem

**Auxiliary fact:** One and only one is true

$$\exists\ x:\ Ax \ge 0,\ Ax \ne 0 \text{ (i)} \quad vs \quad \exists\ y > 0:\ y^T A = 0. \text{ (ii)}$$

**Hint:** (i) $\Rightarrow\ \forall\ y > 0 \to 0 < \langle y, Ax \rangle = \langle A^T y, x \rangle \ne 0\ \Rightarrow$ (ii) is false; (ii) $\Rightarrow$ Convex hull of rows of $A$ is hyperplane or whole space $\Rightarrow$ (i) is false.

$$\neg \exists\ x:\ Ax \ge 0 \left( Ax \ne 0 \right), \quad A = \begin{pmatrix} S \cdot u & C_u \\ S \cdot d & C_d \\ S & C \\ -S & -C \end{pmatrix} \quad \Leftrightarrow \quad C = \frac{1-d}{u-d} C_u + \frac{u-1}{u-d} C_d,$$

where we know $d < 1 < u$, $C_u$, $C_d$.

*Ross S.* An elementary introduction to mathematical finance. Cambridge University Press, 2011.

## Computation of Wasserstein barycenter (M. Cuturi at al.)

$$H_W(L) = \min_{\substack{\sum_{j=1}^{n} x_{ij} = L_i,\ \sum_{i=1}^{n} x_{ij} = W_j \\ x_{ij} \geq 0,\ i,j=1,\dots,n}} \left\{ \gamma \sum_{i,j=1}^{n} x_{ij} \ln x_{ij} + \sum_{i,j=1}^{n} c_{ij} x_{ij} \right\} =$$

$$= \max_{\lambda,\mu} \left\{ \langle \lambda, L \rangle + \langle \mu, W \rangle - \gamma \sum_{i,j=1}^{n} \exp\left( \frac{-c_{ij} + \lambda_i + \mu_j}{\gamma} - 1 \right) \right\} =$$

$$= \max_{\lambda} \left\{ \langle \lambda, L \rangle - \underbrace{\gamma \sum_{j=1}^{n} W_j \ln\left( \frac{1}{W_j} \sum_{i=1}^{n} \exp\left( \frac{-c_{ij} + \lambda_i}{\gamma} \right) \right) \Bigg\},}_{H_W^*(\lambda)} \qquad (*)$$

$$L \in S_n(1) = \left\{ L \geq 0 : \sum_{k=1}^{n} L_k = 1 \right\} \ (W \in S_n(1), \gamma > 0).$$

Due to Demyanov–Danskin theorem for $L \in S_n(1)$ function $H_W(L)$ is smooth with

$$\nabla H_W(L) = \lambda^*,$$

where $\lambda^*$ is unique solution of (*), that satisfy $\langle \lambda^*, e \rangle = 0$. Moreover

$$H_W^*(\lambda) = \max_{L \in S_n(1)} \left\{ \langle \lambda, L \rangle - H_W(L) \right\} = \gamma \sum_{j=1}^{n} W_j \ln\left( \frac{1}{W_j} \sum_{i=1}^{n} \exp\left( \frac{-c_{ij} + \lambda_i}{\gamma} \right) \right).$$

Wasserstein barycenter calculation problem has the following form:

$$\sum_{k=1}^{m} H_{W_k}(L) \to \min_{L \in S_n(1)}. \qquad (**)$$

Unfortunately, $H_{W_k}(L)$ and theirs gradients can't be calculate explicitly.

Let's reformulate (**) in a dual (explicit) manner

$$-\sum_{k=1}^{m} H_{W_k}\left(L_k\right) \to \max_{\substack{L_1 = L_m \mid \lambda^1 \\ \cdots\cdots\cdots \\ L_{m-1} = L_m \mid \lambda^{m-1} \\ L_1, \ldots, L_m \in S_n(1)}},$$

$$\sum_{k=1}^{m-1} \max_{L_k \in S_n(1)} \left\{ \left\langle \lambda^k, L_k \right\rangle - H_{W_k}\left(L_k\right) \right\} + \max_{L_m \in S_n(1)} \left\{ \left\langle -\sum_{k=1}^{m-1} \lambda^k, L_m \right\rangle - H_{W_m}\left(L_m\right) \right\} \to \min_{\lambda^1, \ldots, \lambda^{m-1} \in \mathbb{R}^n},$$

$$\sum_{k=1}^{m-1} H_{W_k}^*\left(\lambda^k\right) + H_{W_m}^*\left(-\sum_{k=1}^{m-1} \lambda^k\right) \to \min_{\lambda^1, \ldots, \lambda^{m-1} \in \mathbb{R}^n}, \qquad (***)$$

$$L_* = \nabla H_{W_k}^*\left(\lambda_*^k\right) \text{ for all } k = 1, \ldots, m-1,$$

where $L_*$ and $\left\{\lambda_*^k\right\}_{k=1}^{m-1}$ – unique solutions of problems (**), (***).

## When it is worth to solve dual problem instead of primal one?

Assume that $Ax = b$ is compatible. Let (conditional optimization problem)

$$\frac{1}{2}\|x\|_2^2 \to \min_{Ax=b}.$$

We can build dual problem (strong duality)

$$\min_{Ax=b} \frac{1}{2}\|x\|_2^2 = \min_x \max_\lambda \left\{ \frac{1}{2}\|x\|_2^2 + \langle b - Ax, \lambda \rangle \right\} =$$

$$= \max_\lambda \min_x \left\{ \frac{1}{2}\|x\|_2^2 + \langle b - Ax, \lambda \rangle \right\} = \max_\lambda \left\{ \langle b, \lambda \rangle - \frac{1}{2}\|A^T\lambda\|_2^2 \right\}.$$

Since $Ax = b$ is compatible, then due to Fredgolm's theorem there is no such $\lambda$, that $A^T\lambda = 0$ and $\langle b, \lambda \rangle > 0$, hence dual problem has finite solution.

Indeed, if there exists such $x$ that $Ax = b$ then for all $\lambda$: $\langle Ax, \lambda \rangle = \langle b, \lambda \rangle$. Hence, $\langle x, A^T \lambda \rangle = \langle b, \lambda \rangle$. Assume that there exists such a $\lambda$, that $A^T \lambda = 0$ and $\langle b, \lambda \rangle > 0$. If it is so we observe a contradiction:

$$0 = \langle x, A^T \lambda \rangle = \langle b, \lambda \rangle > 0.$$

So instead of conditional primal optimization problem one can solve

$$\langle b, \lambda \rangle - \frac{1}{2} \| A^T \lambda \|_2^2 \to \max_{\lambda} \text{ (dual problem – unconditional!)}$$

and reestablish solution of the primal problem from $x(\lambda) = A^T \lambda$.

If matrix $A$ is not a full rank matrix then dual problem have many solutions (affine manifold). But all of them lead to the same (unique) primal solution $x(\lambda) \equiv x_*$.

## When it is worth to solve dual problem instead of primal one?

$$\tilde{F}(x) = \langle c, x \rangle + \|x\|_a^2 + \gamma \sum_{k=1}^{n} x_k \ln x_k \to \min_{x \in S_n(1)}, \text{ // see Lecture 5}$$

or equivalently

$$\langle c, x \rangle + t + \gamma \sum_{k=1}^{n} x_k \ln x_k \to \min_{\substack{x \in S_n(1),\, \|x\|_a^a \le t^{a/2}, \\ 0 \le t \le n^{2/a},\, 0 \le x_k \le 1,\, k=1,\dots,n}} .$$

Using Sion–Kakutani theorem we can build dual problem (strong duality)

$$\tilde{G}(\lambda) = \min_{\substack{0 \le t \le n^{2/a}, \\ 0 \le x_k \le 1,\, k=1,\dots,n}} \left\{ \sum_{k=1}^{n} c_k x_k + t + \lambda_1 \cdot \left( \sum_{k=1}^{n} x_k - 1 \right) + \lambda_2 \cdot \left( \sum_{k=1}^{n} x_k^a - t^{a/2} \right) + \gamma \sum_{k=1}^{n} x_k \ln x_k \right\} \to \max_{\lambda_1 \in \mathbb{R},\, \lambda_2 \ge 0} .$$

We can connect primal variables $(t, x)$ with dual $\lambda = (\lambda_1, \lambda_2)$ $(1 < a < 2)$:

$$t(\lambda) = \min\left\{ \left( \frac{\lambda_2 a}{2} \right)^{\frac{2}{2-a}}, n^{\frac{2}{a}} \right\},$$

and to find $x(\lambda)$ we have to solve $n$ one-dimensional convex optimization sub-problems on finite line segments. So we can obtain $x(\lambda)$ with precision $\varepsilon$ after $\mathrm{O}\left(n\ln\left(n/\varepsilon\right)\right)$ arithmetic operation using half partition line segment method. Using Demyanov–Danskin's formula one can obtain

$$\frac{\partial \tilde{G}}{\partial \lambda_1} = \sum_{k=1}^{n} x_k(\lambda) - 1, \quad \frac{\partial \tilde{G}}{\partial \lambda_2} = \sum_{k=1}^{n} x_k(\lambda)^a - t(\lambda)^{a/2}.$$

We'd like to solve dual problem with ellipsoid method. Since that we have to bound somehow the dual variables. We use Slater's approach

$$-\|c\|_\infty - \gamma \ln n \le \tilde{F}_* = \tilde{G}_* \le \sum_{k=1}^{n} c_k \bar{x}_k + \bar{t} + \lambda_1 \cdot \left( \sum_{k=1}^{n} \bar{x}_k - 1 \right) +$$

$$+\lambda_2 \cdot \left( \sum_{k=1}^{n} \bar{x}_k^a - \bar{t}^{\,a/2} \right) + \gamma \sum_{k=1}^{n} \bar{x}_k \ln \bar{x}_k .$$

If $\lambda_1 \geq 0$, then put $\bar{t} = 1,\ \bar{x}_k = 1/(2n),\ k = 1,\ldots,n$. Hence

$$\frac{1}{2}\lambda_1 + \frac{1}{2}\lambda_2 \leq 2\|c\|_\infty + 2\gamma\ln(n) + 1,$$

If $\lambda_1 < 0$, then put $\bar{t} = 8,\ \bar{x}_k = 2/n,\ k = 1,\ldots,n$. Hence

$$|\lambda_1| + \frac{1}{2}\lambda_2 \leq 3\|c\|_\infty + 2\gamma\ln(2n) + 8.$$

Anyway, we have

$$\|\lambda_*\|_1 \leq 6\|c\|_\infty + 4\gamma\ln(2n) + 16 \overset{def}{=} C.$$

So we can restrict ourselves by solving bounded dual problem

$$-\tilde{G}(\lambda) \to \min_{\lambda_1 \in \mathbb{R},\ \lambda_2 \geq 0,\ \|\lambda\|_1 \leq C}.$$

Using ellipsoid method (Lecture 3) we can find $\varepsilon$-solution of this problem after $\mathrm{O}\left(r^2\ln(C/\varepsilon)\right)$ iterations, the cost of one iteration $\mathrm{O}\left(r^2 + n\ln(n/\varepsilon)\right)$, $r = 2$.

To be continued...