Convex Optimization for Data Science

Gasnikov Alexander

gasnikov.av@mipt.ru

Lecture 5. Primal-duality, regularization, restarts technique, mini-batch & Inexact oracle. Universal methods

November, 2016

Main books:

Nemirovski A. Efficient methods in convex programming. Technion, 1995. http://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf

Nesterov Yu. Introduction Lectures on Convex Optimization. A Basic Course. Applied Optimization. – Springer, 2004.

Nemirovski A. Lectures on modern convex optimization analysis, algorithms, and engineering applications. – Philadelphia: SIAM, 2013.

Devolder O. Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization: PhD thesis. – CORE UCL, March 2013.

Bubeck S. Convex optimization: algorithms and complexity // In Foundations and Trends in Machine Learning. – 2015. – V. 8. – no. 3-4. – P. 231–357.

Gasnikov A.V. Searching equilibriums in large transport networks. Doctoral Thesis. MIPT, 2016. https://arxiv.org/ftp/arxiv/papers/1607/1607.03142.pdf

Structure of Lecture 5

- Basic estimations
- Universal Similar Triangles Method
- Optimal estimation for convex optimization problems
- Mini-batch’ing. Stochastic oracle
- Inexact oracle (Devolder–Glineur–Nesterov)
- Min Max problem
- Min Min problem
- Strongly convex composite
- Regularization technique
- Restarts technique
- Primal-dual methods

Basic estimations

$$F(x) = f(x) + h(x) \to \min_{x \in Q}.$$

We assume that

$$E\left[F\left(x^N\right)\right] - F_* \le \varepsilon.$$

N – number of required iterations: calculations of (stochastic) gradient f.

R – "distance" between starting point and the nearest solution.

N	$E\left[\left\|\partial_x f(x,\xi)\right\|_*^2\right] \le M^2$	$\left\|\nabla f(y) - \nabla f(x)\right\|_* \le L\left\|y - x\right\|$	$E\left[\left\|\nabla_x f(x,\xi) - \nabla f(x)\right\|_*^2\right] \le D$
$F(x)$ convex	$\frac{M^2R^2}{\varepsilon^2}$	$\sqrt{\frac{LR^2}{\varepsilon}}$	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \frac{DR^2}{\varepsilon^2}\right\}$
$F(x)$ μ-strongly convex in $\left\| \right\|$	$\frac{M^2}{\mu\varepsilon}$	$\sqrt{\frac{L}{\mu}}\left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil$	$\max\left\{\sqrt{\frac{L}{\mu}}\left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil, \frac{D}{\mu\varepsilon}\right\}$

If norm is non euclidian then the last row is true up to $O(\ln n)$-factor.

Universal method (Yu. Nesterov, 2013)

We consider composite convex optimization problem

$$F(x) = f(x) + h(x) \to \min_{x \in Q}. \quad (1)$$

Where $R^2 = V\left(x_*, y^0\right)$, and

$$V(x,z) = d(x) - d(z) - \left\langle \nabla d(z), x - z \right\rangle;$$

$d(x) \geq 0$ ($d\left(y^0\right) = 0$, $\nabla d\left(y^0\right) = 0$) is strongly convex in norm $\| \;\|$ with constant ≥ 1; x_* – is the solution of (1) (if the solution is not unique than we can choose such a solution x_* that minimize $V\left(x_*, y^0\right)$).

Assumption 1. *Let*

$$\left\|\nabla f(y) - \nabla f(x)\right\|_* \leq L_\nu \left\|y - x\right\|^\nu, \; \nu \in [0,1].$$

Assumption 2. *Let* $f(x)$ – μ*-strongly convex function in norm* $\| \;\|$, *i.e. for arbitrary* $x, y \in Q$ *holds*

$$f(y)+\langle \nabla f(y), x-y\rangle + \frac{\mu}{2}\|x-y\|^2 \le f(x).$$

Introduce (in euclidian case $\tilde{\omega}_n = 1$)

$$\tilde{\omega}_n = \sup_{x,y\in Q}\frac{2V(x,y)}{\|y-x\|^2},\ \tilde{\mu} = \mu/\tilde{\omega}_n .$$

$$\varphi_0(x) = V(x,y^0) + \alpha_0\left[f(y^0) + \langle \nabla f(y^0), x-y^0\rangle + \tilde{\mu}V(x,y^0) + h(x)\right],$$

$$\varphi_{k+1}(x) = \varphi_k(x) + \alpha_{k+1}\left[f(y^{k+1}) + \langle \nabla f(y^{k+1}), x-y^{k+1}\rangle + \tilde{\mu}V(x,y^k) + h(x)\right].$$

Universal Similar Triangles Method (2016)

Put

$$A_0 = \alpha_0 = 1 / L_0^0,\ k = 0,\ j_0 = 0.$$

Since

$$f\left(x^0\right) > f\left(y^0\right) + \left\langle \nabla f\left(y^0\right), x^0 - y^0 \right\rangle + \frac{L_0^{j_0}}{2}\left\|x^0 - y^0\right\|^2 + \frac{\alpha_0}{2A_0}\varepsilon,$$

where

$$x^0 := u^0 := \arg\min_{x \in Q} \varphi_0\left(x\right),\ \left(A_0 :=\right)\alpha_0 := \frac{1}{L_0^{j_0}},$$

fulfils

$$j_0 := j_0 + 1;\ L_0^{j_0} := 2^{j_0} L_0^0.$$

1. $L_{k+1}^{0} = L_{k}^{j_k} / 2,\ j_{k+1} = 0.$

2. $\alpha_{k+1} := \frac{1 + A_k \tilde{\mu}}{2L_{k+1}^{j_{k+1}}} + \sqrt{\frac{1 + A_k \tilde{\mu}}{4\left(L_{k+1}^{j_{k+1}}\right)^2} + \frac{A_k \cdot (1 + A_k \tilde{\mu})}{L_{k+1}^{j_{k+1}}}},\ A_{k+1} := A_k + \alpha_{k+1},$ (*)

$y^{k+1} := \frac{\alpha_{k+1} u^k + A_k x^k}{A_{k+1}},\ u^{k+1} := \arg\min_{x \in Q} \varphi_{k+1}(x),\ x^{k+1} := \frac{\alpha_{k+1} u^{k+1} + A_k x^k}{A_{k+1}}.$ (**)

Since

$$f\left(y^{k+1}\right) + \left\langle \nabla f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle + \frac{L_{k+1}^{j_{k+1}}}{2}\left\| x^{k+1} - y^{k+1} \right\|^2 + \frac{\alpha_{k+1}}{2A_{k+1}}\varepsilon < f\left(x^{k+1}\right),$$

fulfils $j_{k+1} := j_{k+1} + 1;\ L_{k+1}^{j_{k+1}} = 2^{j_{k+1}} L_{k+1}^{0}$; (*), (**).

3. If stopping rule doesn't satisfy, put $k := k + 1$ and **go to** 1.

Theorem 1. *Let assumption 1 is true for at least* $\nu = 0$ *and assumption 2 fulfils with* $\mu \geq 0$ *(it is possible to take* $\mu = 0$*). Then USTM for (1) converges according to the estimation*

$$F\left(x^N\right) - \min_{x \in Q} F\left(x\right) \leq \varepsilon,$$

$$N\left(\varepsilon\right) \approx \min\left\{ \inf_{\nu \in [0,1]} \left(\frac{L_\nu \cdot \left(16R\right)^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+3\nu}}, \right.$$

$$\left. \inf_{\nu \in [0,1]} \left\{ \left(\frac{8 L_\nu^{\frac{2}{1+\nu}} \tilde{\omega}_n}{\mu \varepsilon^{\frac{1-\nu}{1+\nu}}} \right)^{\frac{1+\nu}{1+3\nu}} \ln^{\frac{2+2\nu}{1+3\nu}} \left(\frac{16 L_\nu^{\frac{4+6\nu}{1+\nu}} R^2}{\left(\mu / \tilde{\omega}_n\right)^{\frac{1+\nu}{1+3\nu}} \varepsilon^{\frac{5+7\nu}{2+6\nu}}} \right) \right\} \right\}. \quad (2)$$

Discussion

At each iteration USTM requires in average for calculations of function f values in ≈ 4 points, and its gradient in ≈ 2 points.

Moreover for $k = 0,1,2,...$ it holds

$$\left\|u^k - x_*\right\|^2 \le 2R^2,\ \max\left\{\left\|x^k - x_*\right\|^2, \left\|y^k - x_*\right\|^2\right\} \le 4R^2 + 2\left\|x^0 - y^0\right\|^2.$$

If inf is attained under $\nu = 0$, then USTM corresponds (up to a logarithmic factor) for the rate of convergence to Mirror Descent, and if inf is attained under $\nu = 1$ then USTM corresponds to STM (see Lecture 3).

Gasnikov A., Nesterov Yu. Universal fast gradient method for stochastic composit optimization problems // Comp. Math. & Math. Phys. 2016. (in print) arXiv:1604.05275

Assume, that instead of real gradients we have only stochastic gradients $\nabla f(x) \to \nabla f(x,\xi)$ (one can generalize in the case when also instead of the function's values we have only its realizations $f(x) \to f(x,\xi)$).

Assumption 3. *Let for all* $x \in Q$

$$E_{\xi}\left[\nabla f(x,\xi)\right] = \nabla f(x) \text{ and } E_{\xi}\left[\left\|\nabla f(x,\xi) - \nabla f(x)\right\|_*^2\right] \le D.$$

Let's introduce (**mini-batch**'ing) $\boxed{\bar{\nabla}^m f(x) = \frac{1}{m}\sum_{k=1}^{m} \nabla f\left(x,\xi^k\right)}$, where ξ^k – i.i.d. (distributed the same as ξ),

$$\varphi_0(x) = V\left(x,y^0\right) + \alpha_0\left[f\left(y^0\right) + \left\langle \bar{\nabla}^m f\left(y^0\right), x - y^0\right\rangle + \tilde{\mu} V\left(x,y^0\right) + h(x)\right],$$

$$\varphi_{k+1}(x) = \varphi_k(x) + \alpha_{k+1}\left[f\left(y^{k+1}\right) + \left\langle \bar{\nabla}^m f\left(y^{k+1}\right), x - y^{k+1}\right\rangle + \tilde{\mu} V\left(x,y^k\right) + h(x)\right].$$

If additionally in theorem 1 assumption 3 is true and if we introduce on the step 2 USTM $m_{k+1} := 8DA_{k+1} \big/ L_{k+1}^{j_{k+1}} \alpha_{k+1} \varepsilon$ and change stopping rule at this step

$$f\left(y^{k+1}\right) + \left\langle \bar{\nabla}^{m_{k+1}} f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle + \frac{L_{k+1}^{j_{k+1}}}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 + \frac{\alpha_{k+1}}{2A_{k+1}} \varepsilon < f\left(x^{k+1}\right),$$

then estimation (2) changes: $N(\varepsilon) \to 2N(\varepsilon/4)$. At each iteration method requires in average for calculations of function f values in ≈ 4 points. One can also obtain the following estimation of total number of stochastic gradients' calculations for the (average) precision 2ε (up to a $\sim \ln n$ factor in non euclidian case, see arXiv:1601.07592, Proposition 6)

$$2 \cdot \min\left\{ \frac{64DR^2}{\varepsilon^2}, \frac{8D\tilde{\omega}_n}{\mu\varepsilon} \ln\left(\frac{8L_0^{j_0} R^2}{\varepsilon} \right) \right\} + 4N(\varepsilon/4). \tag{3}$$

We use Fenchel inequality and the fact that $E[RHS] \le 2D \big/ \left(L_{k+1}^{j_{k+1}} m \right)$ (up to a $\sim \ln n$ factor):

$$\left\langle \bar{\nabla}^{m_{k+1}} f\left(y^{k+1}\right) - \nabla f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle - \frac{L_{k+1}^{j_{k+1}} / 2}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 \le \frac{2}{L_{k+1}^{j_{k+1}}} \left\| \bar{\nabla}^{m_{k+1}} f\left(y^{k+1}\right) - \nabla f\left(y^{k+1}\right) \right\|_*^2.$$

Estimations (2), (3) save their view, if we work with inexact (δ, L, μ)-oracle (Devolder–Glineur–Nesterov, 2011) with

$$\delta = \mathrm{O}\left(\varepsilon / N(\varepsilon)\right) \text{ and } L = \mathrm{O}\left(\max_{k=0,\ldots,N} L_k^{j_k}\right).$$

This oracle on request, determines by only one point x, returns such a pair $\left(f_\delta(x), g_\delta(x,\xi)\right)$ (one can generalize for the case $f_\delta(x) \to f_\delta(x,\xi)$), that for all $x \in Q \to E_\xi\left[\left\|g_\delta(x,\xi) - E_\xi\left[g_\delta(x,\xi)\right]\right\|_*^2\right] \le D$ and for all $x, y \in Q$

$$\frac{\mu}{2}\|y-x\|^2 \le f(y) - f_\delta(x) - \left\langle E_\xi\left[g_\delta(x,\xi)\right], y-x\right\rangle \le \frac{L}{2}\|y-x\|^2 + \delta.$$

All the bounds mentioned above are optimal up to a logarithmic factor (A. Nemirovski, 1979, Devolder–Glineur–Nesterov, 2011, P. Dvurechensky, 2014).

Idea behind the Universal method

From

$$\left\|\nabla f\left(y\right)-\nabla f\left(x\right)\right\|_{*}\le L_{\nu}\left\|y-x\right\|^{\nu},\nu\in\left[0,1\right]$$

one has

$$0\le f\left(y\right)-f\left(x\right)-\left\langle\nabla f\left(x\right),y-x\right\rangle\le\frac{L}{2}\left\|y-x\right\|^{2}+\delta,\; L=L_{\nu}\cdot\left[\frac{L_{\nu}}{2\delta}\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}.$$

Since for arbitrary fast gradient method with inexact oracle

$$N^{2}\sim\frac{LR^{2}}{\varepsilon},\; L\sim L_{\nu}\cdot\left(\frac{L_{\nu}}{\delta}\right)^{\frac{1-\nu}{1+\nu}},\;\delta\sim\frac{\varepsilon}{N},$$

we have

$$N^{2}\sim\frac{L_{\nu}^{\frac{2}{1+\nu}}R^{2}}{\varepsilon\delta^{\frac{1-\nu}{1+\nu}}}\sim\frac{L_{\nu}^{\frac{2}{1+\nu}}R^{2}}{\varepsilon^{\frac{2}{1+\nu}}N^{-\frac{1-\nu}{1+\nu}}}\Rightarrow N^{\frac{1+3\nu}{1+\nu}}\sim\frac{L_{\nu}^{\frac{2}{1+\nu}}R^{2}}{\varepsilon^{\frac{2}{1+\nu}}}\Rightarrow N\sim\left(\frac{L_{\nu}R^{1+\nu}}{\varepsilon}\right)^{\frac{2}{1+3\nu}}.$$

Non accelerated methods

For gradient descent and conditional gradient descent (see Lecture 3)

$$f\left(x^N\right) - f_* = \mathrm{O}\left(\frac{LR^2}{N} + \delta R\right) \text{ // } f\left(x^N\right) - f_* = \mathrm{O}\left(\frac{MR}{\sqrt{N}} + \delta R\right) \text{ for MD.}$$

If one chooses in smooth-methods stochastic gradients $\nabla f\left(x, \xi^k\right)$ with variance D and uses mini-batches $\bar{\nabla}^m f\left(x\right)$ with proper m, then one can obtain the following analogues of formulas from the table above

$$f\left(x^N\right) - f_* = \tilde{\mathrm{O}}\left(\max\left\{\frac{LR^2}{N}, \sqrt{\frac{DR^2}{N}}\right\}\right). \text{ // for STM } \tilde{\mathrm{O}}\left(\max\left\{\sqrt{\frac{LR^2}{N}}, \sqrt{\frac{DR^2}{N}}\right\}\right)$$

One can generalize for strongly convex case and also generalize CGD, GD (its universal variant) for non convex case (S. Ghadimi, G. Lan, E. Hazan e.t.c.).

Illustrative Examples

Let's consider concrete examples. In all these examples we assume that $L = L_1 < \infty$ (see denotation in assumption 1 above).

Example 1 (min max problem). Let's consider saddle-point problem (Fenchel's type functional)

$$f(x) = \max_{\|y\|_2 \le R_y} \left\{ G(y) + \langle By, x \rangle \right\} \to \min_{\|x\|_2 \le R_x},$$

where $G(y)$ – is μ-strongly concave in 2-norm with Lipschitz constant of gradient L_G in 2-norm. Then $f(x)$ is smooth, with Lipschitz constant of gradient in 2-norm $L_f = \sigma_{\max}(B)/\mu$. It seems that one can minimize $f(x)$ for $O\left(\sqrt{\sigma_{\max}(B) R_x^2 / (\mu\varepsilon)}\right)$ iteration, where ε – is desirable precision on functional convergence. But this estimation is true if we can exactly calculate $\nabla f(x) = By^*(x)$ (Demyanov–Danskin's formula, see Lecture 1), where $y^*(x)$ – is the solution of inner problem for y (under fixed x). In reality we can solve this inner problem only numerically (that is with some error). If we solve inner problem by (U)STM with precision $\delta/2$ (for that we have to do $\mathrm{O}\left(\sqrt{L_G/\mu}\ln\left(L_G R_y^2/\delta\right)\right)$ iterations), then

$$\left(G\left(y_{\delta/2}(x)\right) + \left\langle By_{\delta/2}(x), x\right\rangle, By_{\delta/2}(x)\right),$$

where $y_{\delta/2}(x)$ – is $\delta/2$-solution of inner problem, is $(\delta, 2L_f, 0)$-oracle (Devolder–Glineur–Nesterov, 2013). By choosing $\delta = \mathrm{O}\left(\varepsilon\sqrt{\varepsilon/\left(L_f R_x^2\right)}\right)$, one can obtain after

$$\mathrm{O}\left(\sqrt{\frac{L_G\sigma_{\max}\left(B\right)R_x^2}{\mu^2\varepsilon}}\ln\left(\frac{L_f L_G R_x^2 R_y^2}{\varepsilon}\right)\right)$$

iterations (at each iteration matrix B is multiplied on vector one calculates gradient of $G\left(y\right)$) ε-solution of initial problem of minimization of $f\left(x\right)$. Note that if $G\left(y\right)$ isn't strongly convex, then for finding such $\left(x^N, y^N\right)$ that (this is almost the same as to solve initial problem with precision ε)

$$\max_{\|y\|_2\le R_y}\left\{G\left(y\right)+\left\langle By, x^N\right\rangle\right\}-\min_{\|x\|_2\le R_x}\left\{G\left(y^N\right)+\left\langle By^N, x\right\rangle\right\}\le\varepsilon,$$

one have to do at least $\Omega\left(\max\left\{L_G R_y^2, \sigma_{\max}\left(B\right)R_x R_y\right\}/\varepsilon\right)$ iterations. □

Example 2 (min min problem). Let $f(x)=\min_{y\in Q}\Phi(y,x)$, where Q – is compact convex set and $\Phi(y,x)$ – is such smooth convex function that

$$\left\|\nabla\Phi(y',x')-\nabla\Phi(y,x)\right\|_2 \le L\left\|(y',x')-(y,x)\right\|_2, \text{ for all } y,y'\in Q.$$

Assume that for all x (for simplicity we consider $x\in\mathbb{R}^n$) one can find such $\tilde{y}=\tilde{y}(x)\in Q$ that

$$\max_{z\in Q}\left\langle\nabla_y\Phi(\tilde{y},x),\tilde{y}-z\right\rangle\le\delta.$$

Then

$$\Phi(\tilde{y},x)-f(x)\le\delta,\ \left\|\nabla f(x')-\nabla f(x)\right\|_2\le L\left\|x'-x\right\|_2,$$

and $\left(\Phi(\tilde{y},x)-2\delta,\nabla_y\Phi(\tilde{y},x)\right)$ is $(6\delta,2L,0)$-oracle for $f(x)$. □

Example 3 (see Lecture 2). Let

$$F(x) = \frac{1}{2}\|Ax - b\|_2^2 + \mu \sum_{k=1}^{n} x_k \ln x_k \to \min_{\sum_{k=1}^{n} x_k = 1,\ x \geq 0}.$$

We'll consider two cases **a)** $0 < \mu \ll \varepsilon/(2\ln n)$; **b)** $\mu \gg \varepsilon/(2\ln n)$.

a) We choose $\|\ \| = \|\ \|_1$. Put

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2,\ h(x) = \mu \sum_{k=1}^{n} x_k \ln x_k,\ Q = S_n(1) = \left\{ x \geq 0 : \sum_{k=1}^{n} x_k = 1 \right\},$$

$L = \max_{k=1,\dots,n} \left\| A^{\langle k \rangle} \right\|_2^2$, where $A^{\langle k \rangle}$ – k-th column of A. For the case a) one can choose $d(x) = \ln n + \sum_{k=1}^{n} x_k \ln x_k$. Then $V(x,z) = \sum_{k=1}^{n} x_k \ln(x_k / z_k)$, $R^2 \leq \ln n$.

Here we have such a situation when Bregman's divergence $V(x,z)$ coincides in form with composite. Since that we have explicit formulas for iteration step of (U)STM method. Therefore the cost of one iteration is $\mathrm{O}(nnz(A))$, where $nnz(A)$ – is number of non-zero elements of A (we assume that $nnz(A) \ge n$). The total number of required iterations is the following (see Lecture 3)

$$N = \mathrm{O}\left(\sqrt{\frac{\max\limits_{k=1,\dots,n}\left\|A^{\langle k\rangle}\right\|_2^2 \ln n}{\varepsilon}}\right). \ \square$$

Unfortunately, it isn't good to use (U)STM directly for the case b) since $f(x)$ isn't strongly convex. But one can built a proper method from (U)STM by **restarts technique**. But we start with **regularization technique**.

Let's introduce μ-strongly convex in norm $\| \ \|$ problem ($\mu > 0$)

$$F^{\mu}\left(x\right) = F\left(x\right) + \mu V\left(x, y^0\right) \to \min_{x \in Q}. \tag{4}$$

Let F_*^{μ} – is optimal value in (4) and F_* – is optimal value in (1).

Proposition 1 (regularization). *Let*

$$\mu \le \frac{\varepsilon}{2V\left(x_*, y^0\right)} = \frac{\varepsilon}{2R^2},$$

and there exists such $x^N \in Q$ *that*

$$F^{\mu}\left(x^N\right) - F_*^{\mu} \le \varepsilon/2.$$

Then

$$F\left(x^N\right) - F_* \le \varepsilon.$$

Vasiliev F.P. Optimization methods. MCCME, 2011. [in Russia]

Proposition 2 (restarts). *Let assumption 1 is true with* $\nu = 1$ *(*$L = L_1$*), function* $F(x)$ *– is* μ*-strongly convex in norm* $\| \ \|$*. Let* $x^{\bar{N}}(y^0)$ *– is return of STM (or USTM with* $\mu = 0$*), with starting point* y^0*, after*

$$\bar{N} = \sqrt{\frac{16L\omega_n}{\mu}}$$

iterations, where (one should compare with $\tilde{\omega}_n$ *introduced above)*

$$\omega_n = \sup_{x \in Q} \frac{2V(x, y^0)}{\|x - y^0\|^2}.$$

Put $\left[x^{\bar{N}}(y^0)\right]^1 = x^{\bar{N}}(y^0)$ *and determine for induction*

$$\left[x^{\bar{N}}(y^0)\right]^{k+1} = x^{\bar{N}}\left(\left[x^{\bar{N}}(y^0)\right]^k\right),\ k = 1, 2, \ldots.$$

Note that on $(k+1)$*-th restart we redeterminate prox-function*

$$d^{k+1}(x) = d\left(x - \left[x^{\bar{N}}\left(y^0\right)\right]^k + y^0\right) \geq 0,$$

For the following is true

$$d^{k+1}\left(\left[x^{\bar{N}}\left(y^0\right)\right]^k\right) = 0,\ \nabla d^{k+1}\left(\left[x^{\bar{N}}\left(y^0\right)\right]^k\right) = 0.$$

Then
$$F\left(\left[x^{\bar{N}}\left(y^0\right)\right]^k\right) - F_* \leq \frac{\mu\left\|y^0 - x_*\right\|^2}{2^{k+1}}.$$

Dvurechensky–Kamzolov proposes restart technique for Intermediate Universal Method.

These two techniques generate optimal methods from the optimal ones. Problem of regularization technique: requires R. Problem of restarts technique: requires μ. Important open problem: Propose universal method in μ.

For more details see: arXiv:1204.3982; arXiv:1609.07358; arXiv:1702.03828

Regularization technique $\mu \sim \varepsilon / R^2$

N	$E\left[\left\lVert \partial_x f(x,\xi)\right\rVert_*^2\right] \le M^2$	$\left\lVert \nabla f(y) - \nabla f(x)\right\rVert_* \le L\left\lVert y-x\right\rVert$	$E\left[\left\lVert \nabla_x f(x,\xi) - \nabla f(x)\right\rVert_*^2\right] \le D$
$F(x)$ μ-strongly convex in $\lVert\ \rVert$	$\frac{M^2}{\mu\varepsilon}$	$\sqrt{\frac{L}{\mu}}\left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil$	$\max\left\{\sqrt{\frac{L}{\mu}}\left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil, \frac{D}{\mu\varepsilon}\right\}$
$F(x)$ convex	$\frac{M^2 R^2}{\varepsilon^2}$	$\sqrt{\frac{LR^2}{\varepsilon}}$	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \frac{DR^2}{\varepsilon^2}\right\}$

Restarts technique (inverse to regularization)

N	$E\left[\left\lVert \partial_x f(x,\xi)\right\rVert_*^2\right] \le M^2$	$\left\lVert \nabla f(y) - \nabla f(x)\right\rVert_* \le L\left\lVert y-x\right\rVert$	$E\left[\left\lVert \nabla_x f(x,\xi) - \nabla f(x)\right\rVert_*^2\right] \le D$
$F(x)$ convex	$\frac{M^2 R^2}{\varepsilon^2}$	$\sqrt{\frac{LR^2}{\varepsilon}}$	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \frac{DR^2}{\varepsilon^2}\right\}$
$F(x)$ μ-strongly convex in $\lVert\ \rVert$	$\frac{M^2}{\mu\varepsilon}$	$\sqrt{\frac{L}{\mu}}\left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil$	$\max\left\{\sqrt{\frac{L}{\mu}}\left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil, \frac{D}{\mu\varepsilon}\right\}$

Example 3. b) In this case it's worth to use restarts technique (proposition 2). Unfortunately, for entropy prox function $\omega_n = \infty$. Let's introduce

$$d(x) = \frac{1}{2(a-1)}\|x\|_a^2,\ a = \frac{2\ln n}{2\ln n - 1}.$$

In this case $R^2 = \mathrm{O}(\ln n)$, $\omega_n = \mathrm{O}(\ln n)$. Complexity of one iteration (additional for calculation of gradient – $\mathrm{O}(nnz(A))$) is determine how efficiently one can solve the following problem (see Lecture 1)

$$\tilde{F}(x) = \langle c, x\rangle + \|x\|_a^2 + \bar{\mu}\sum_{k=1}^{n} x_k \ln x_k \to \min_{x\in S_n(1)}.$$

As we've already known the complexity is $\mathrm{O}\left(n\ln^2\left(n/\varepsilon\right)\right)$. This complexity is typically much smaller then $\mathrm{O}\left(nnz\left(A\right)\right)$.

The number of required iterations (see Lecture 3; Theorem 1 and Proposition 2)

$$N = \mathrm{O}\left(\sqrt{\frac{\max\limits_{k=1,\dots,n}\left\|A^{\langle k\rangle}\right\|_2^2 \ln n}{\mu}}\left\lceil \log_2\left(\frac{\mu}{\varepsilon}\right)\right\rceil\right).$$

Note, that from this estimation and proposition 1 one can obtain estimation of example 3 a) (up to $\sim\sqrt{\ln n}$). □

Example 4. (Lyapunov's type optimal control problem). arXiv:1703.00267

$$F\left(u(\cdot)\right)=\int_0^T f^0\left(t,x(t),u(t)\right)dt+\Phi\left(x(T)\right)\to \min_{u(\cdot)\in U\subseteq L_2[0,T]},$$

$$\frac{dx}{dt}=f\left(t,x(t),u(t)\right),\quad x(0)=x^0. \qquad (*)$$

where U is convex, all functions are smooth enough and linear with coefficients depend only on t. This problem is convex!

$$\nabla F\left(u(\cdot)\right)=\left.\frac{\partial H\left(t,x,u,\psi\right)}{\partial u}\right|_{x=x(t,u),\,u=u(t),\,\psi=\psi(t,u)},\ H=f^0+\langle\psi,f\rangle,$$

here $x(t,u)$ is solution of (*) and $\psi(t,u)$ is solution of

$$\frac{d\psi}{dt}=-\frac{\partial H\left(t,x,u,\psi\right)}{\partial x},\quad \psi(T)=\nabla\Phi\left(x(T,u)\right). \qquad (**)$$

Unfortunately, one can't calculate precisely gradient since one should solve two system of ordinary differential equations (*), (**). But one can solve these two systems by introducing the same lattice in t (with the size of each element h: $t^{k+1} - t^k \equiv h$) for both of the systems (*), (**):

$$\frac{x\left(t^{k+1}\right) - x\left(t^k\right)}{h} = f\left(t^k, x\left(t^k\right), u\left(t^k\right)\right), \quad x\left(t^0\right) = x(0) = x^0,$$

$$\frac{\psi\left(t^k\right) - \psi\left(t^{k+1}\right)}{h} = \frac{\partial H}{\partial x}\left(t^{k+1}, x\left(t^{k+1}\right), u\left(t^{k+1}\right), \psi\left(t^{k+1}\right)\right), \quad \psi(T) = \nabla \Phi\left(x\left(t^{T/h}\right)\right).$$

Here we use the standard Euler's scheme with the quality of approximation $\delta \sim he^{cT}$ and the complexity $\sim h^{-1}$. So using the theory above (USTM) one can build a fast gradient descent method with proper choice of $h \sim \varepsilon^{3/2}$. The total complexity $\sim \varepsilon^{-2}$. The same result (about total complexity) is true for (U)GD. But the last method works also with non convex problems (local extreme).

Note, that due to linearity on x:

$$\frac{\partial H(t,x,u,\psi)}{\partial x} \equiv h_0(t) + h_1(t)\psi .$$

Since that instead of Euler's scheme one can use Runge–Kutta's schemes of order $k \ge 2$. Moreover, one can dip (U)GM and (U)STM in one parametric family of intermediate methods (Devolder–Glineur–Nesterov, 2013; P. Dvurechensky, 2014; D. Kamzolov, 2016)

$$F(x^N) - F_* \le \varepsilon,\ N = \mathrm{O}\left(\inf_{\nu\in[0,1]} \left(\frac{L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+2p\nu+\nu}} \right),\ \delta \le \mathrm{O}\left(\frac{\varepsilon}{N^p} \right),\ p \in [0,1].\ //\ \nu = 1$$

The cost of one iteration is still $\mathrm{O}(h^{-1})$, $N \sim \varepsilon^{-1/(1+p)}$, $h^k \sim \delta \sim \varepsilon / N^p \sim \varepsilon^{2-1/(1+p)}$.

Hence, Total complexity $\sim \varepsilon^{-(2/k+(1-1/k)/(1+p))}$. For $k \ge 2$ optimal $p = 1$. □

Primal-duality of STM & USTM

We have to solve the following convex optimization problem

$$g(x) \to \min_{Ax=b,\, x\in Q}, \tag{5}$$

where $g(x)$ is 1-strongly convex function in p-norm $(1 \le p \le 2)$. We build dual problem

$$f(y) = \max_{x\in Q}\left\{\langle y, b - Ax\rangle - g(x)\right\} = \langle y, b - Ax(y)\rangle - g(x(y)) \to \min_{y}. \tag{6}$$

In many applications the main contribution in computational complexity of one iteration gives calculations of Ax, $A^T y$.

Nesterov Yu. Primal-dual subgradient methods for convex problems // Math. Program. Ser. B. – 2009. – V. – 120(1). – P. 261–283.
Nemirovski A., Onn S., Rothblum U.G. Accuracy certificates for computational problems with convex structure // Mathematics of Operation Research. – 2010. – V. 35. – № 1. – P. 52–78.

Let (U)STM with $\| \ \| = \| \ \|_2$, $d(y) = \frac{1}{2}\|y\|_2^2$, $y^0 = 0$, for the problem (5) generates points $\{y^k\}$ (based on these points we build $\varphi_k(y)$), and $\tilde{y}^N$ (in theorem 1 we denote this point x^N). Put

$$x^N = \sum_{k=0}^{N} \lambda_k x(y^k), \ \lambda_k = \alpha_k / A_N \ .$$

Since (x_* – solution of (5))

$$g(x^N) - g(x_*) \leq f(\tilde{y}^N) + g(x^N),$$

the next theorem allows us to calculate the solution of (5) with prescribed precision.

Note: Indeed, all mentioned above method (expect GD) are primal-dual.

Theorem 2. *Let we want to solve problem (5) by passing to the dual problem (6), according to the formulas mentioned above. Let's choose the following stopping rule for (U)STM*

$$f\left(\tilde{y}^N\right)+g\left(x^N\right)\le\varepsilon,\ \left\|Ax^N-b\right\|_2\le\tilde{\varepsilon}.$$

Then (U)STM is stop by making no more than ($L=\max_{\|x\|_p\le 1}\|Ax\|_2^2$*)*

$$6\cdot\max\left\{\sqrt{\frac{LR^2}{\varepsilon}},\sqrt{\frac{LR}{\tilde{\varepsilon}}}\right\}$$

iterations, where $R^2=\|y_*\|_2^2$, y_* *– solution of the problem (6) (if the solution is not unique than we can choose such a solution* y_* *that minimize* R^2*).*

https://arxiv.org/ftp/arxiv/papers/1602/1602.01686.pdf

Primal-duality via regularization

Idea: regularize dual problem (6) (we use $x^N = x\left(y^N\right)$ for solution of (5))

$$f^{\mu}(y) = f(y) + \frac{\mu}{2}\|y\|_2^2 \to \min_y, \ \mu \simeq \varepsilon / \left(2R^2\right). \text{ // we restart on } \mu$$

$$\frac{1}{2L}\left\|\nabla f^{\mu}(y)\right\|_2^2 \le f^{\mu}(y) - f_*^{\mu} \le \frac{1}{2\mu}\left\|\nabla f^{\mu}(y)\right\|_2^2,$$

$$g\left(x(y)\right) - g\left(x_*\right) \le \|y\|_2 \left\|Ax(y) - b\right\|_2.$$

We use **stopping rule**: $\left\|y^N\right\|_2 \left\|Ax\left(y^N\right) - b\right\|_2 \le \varepsilon, \ \left\|Ax\left(y^N\right) - b\right\|_2 \le \tilde{\varepsilon}.$

Oracle calls: $N \simeq \sqrt{\dfrac{2L \cdot \left(\varepsilon + 2R\tilde{\varepsilon}\right)}{\tilde{\varepsilon}^2}} \ln\left(\dfrac{4L \max\limits_{x,y\in Q}\left|g(x) - g(y)\right| \cdot \left(\varepsilon + 2R\tilde{\varepsilon}\right)}{\varepsilon \cdot \tilde{\varepsilon}^2}\right).$

https://arxiv.org/ftp/arxiv/papers/1410/1410.7719.pdf

Convergence on gradient (non strongly convex case)

The structure of the dual functional allows one to obtain $\left\|\nabla f\left(y^N\right)\right\|_2 \sim N^{-2}$.

But in general (without primal-dual structure of f) one can only guarantee

$$\left\|\nabla f\left(y^N\right)\right\|_2 \sim \left(\ln N\right)^2 \big/ N^2 . \text{ // use regularization}$$

In non convex case optimal estimation is

$$\left\|\nabla f\left(y^N\right)\right\|_2 \sim 1\big/\sqrt{N} . \text{ // } \frac{1}{2L}\left\|\nabla f\left(y^N\right)\right\|_2^2 \le f\left(y^N\right) - f_* = \mathrm{O}\left(\frac{LR^2}{N}\right)$$

In general one should use here gradient mapping instead of gradient.

Nesterov Yu. How to make the gradients small // OPTIMA 88. 2012. P. 10–11.
Carmon Y., Duchi J.C., Hinder O., Sidford A. arXiv:1611.00756
Agarwal N., Allen-Zhu Z., Bullins B., Hazan E., Ma T. arXiv:1611.01146

Google problem

$$Ax = \begin{pmatrix} \left(P^T - I\right) \\ 1\ldots\ldots1 \end{pmatrix} x = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = b. \ // \ x \in \mathbb{R}^n,\ n \gg 1 \text{ (Lecture 2)}$$

According to the Frobenius–Perron's theory if matrix P is irreducible then this system has a unique solution (and $x > 0$). Let's reformulate the problem as convex optimization problem

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

One can built a dual problem (Lecture 3)

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

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

Since $Ax = b$ is compatible then for Fredgolm's theorem it's no possible that there exists such λ: $A^T \lambda = 0$ and $\langle b, \lambda \rangle > 0$. Hence the dual problem is solvable (but solution isn't unique). Let's denote λ^* to be the solution of the dual problem

$$\langle b, \lambda \rangle - \frac{1}{2} \|A^T \lambda\|_2^2 \to \max_{\lambda}$$

with minimal 2-norm. Let's introduce (from optimality condition for x): $x(\lambda) = A^T \lambda$. Using (U)STM for the dual problem one can find (Theorem 2)

$$\|Ax^N - b\|_2 = O\left(\frac{L_y R_y}{N^2} \right),$$

where x^N is a convex combination of

$$\left\{x\left(\lambda^k\right)\right\}_{k=1}^N, \quad L_y = \sigma_{\max}\left(A^T\right) = \sigma_{\max}\left(A\right), \; R_y = \left\|\lambda_*\right\|_2.$$

The other way to find Page Rank vector is to solve the system $Ax = b$ or to solve convex optimization problem

$$\frac{1}{2}\left\|Ax - b\right\|_2^2 \to \min_x.$$

Using (U)STM one can obtain ($L_x = \sigma_{\max}\left(A\right) = L_y, \; R_x = \left\|x_*\right\|_2 \le 1$)

$$\left\|Ax^N - b\right\|_2 = \mathrm{O}\left(\frac{\sqrt{L_x}R_x}{N}\right).$$

This is a lower bound for $Ax = b$ for $N \le n$ (Nemirovski–Yudin, 1979). There is no contradiction here, since this $L_y R_y \ll n\sqrt{L_x}R_x$ isn't always true.

Primal-dual method for searching traffic assignment (Lecture 2)

$$\Psi(f) = \sum_{e\in E} \sigma_e(f_e) \to \min_{\substack{f=\Theta x,\, x\in X \\ f_e \le \bar{f}_e,\, e\in E'}}, \ \sigma_e(f_e) = f_e \bar{t}_e, \ e \in E',$$

$$\Upsilon(t) = \underbrace{-\sum_{w\in OD} d_w T_w(t)}_{F(t)} + \sum_{e\in E} \sigma_e^*(t_e) \to \min_{\substack{t_e \ge \bar{t}_e,\, e\in E' \\ t_e \in \operatorname{dom}\sigma_e^*(t_e),\, e\in E\setminus E'}}. \text{ // dual problem}$$

where $T_w(t) = \min_{p\in P_w} \sum_{e\in E} \delta_{ep} t_e$ – the length of the shortest path from i to j ($w = (i, j) \in OD$) on the transport graph weighted by $t = \{t_e\}_{e\in E}$. One can solve find f from the solution of dual problem: $f_e = \bar{f}_e - s_e$, $e \in E'$, where $s_e \ge 0$ – Lagrange's multiplier to $t_e \ge \bar{t}_e$; $\tau_e(f_e) = t_e$, $e \in E \setminus E'$. Note, that for the edge $e \in E'$: $\sigma_e^*(t_e) = \bar{f}_e \cdot (t_e - \bar{t}_e)$ and for $e \in E$ (typically $\mu = 1/4$)

$$\tau_e(f_e) = \bar{t}_e \cdot \left(1 + \gamma \cdot \left(f_e / \bar{f}_e\right)^{\frac{1}{\mu}}\right) \Rightarrow \sigma_e^*(t_e) = \bar{f}_e \cdot \left(\frac{t_e - \bar{t}_e}{\bar{t}_e \cdot \gamma}\right)^{\mu} \frac{(t_e - \bar{t}_e)}{1+\mu}.$$

$$\boxed{t^{k+1} = \arg \min_{\substack{t_e \ge \bar{t}_e,\, e \in E' \\ t_e \in \operatorname{dom} \sigma_e^*(t_e),\, e \in E \setminus E'}} \left\{ \gamma_k \left\{ \left\langle \partial F\left(t^k\right), t - t^k \right\rangle + \sum_{e \in E} \sigma_e^*\left(t_e\right) \right\} + \frac{1}{2}\left\| t - t^k \right\|_2^2 \right\}},$$

where (we use composite Mirror Descent, see Lecture 3)

$$\gamma_k = \varepsilon / M_k^2,\ M_k = \left\| \partial F\left(t^k\right) \right\|_2,$$

Let's introduce
$$\bar{t}^N = \frac{1}{S_N} \sum_{k=0}^{N} \gamma_k t^k,\ S_N = \sum_{k=0}^{N} \gamma_k,$$

$$\boxed{f_e^k \in -\partial_e F\left(t^k\right),\ \bar{f}_e^N = \frac{1}{S_N} \sum_{k=0}^{N} \gamma_k f_e^k,\ e \in E \setminus E';\ \bar{f}_e^N = \bar{f}_e - s_e^N,\ e \in E',}$$

where s_e^N – Lagrange's multiplier to $t_e \ge \bar{t}_e$ in the problem

$$\frac{1}{S_N}\left\{ \sum_{k=0}^{N} \gamma_k \cdot \left\{ \sum_{e \in E'} \partial_e F\left(t^k\right) \cdot \left(t_e - t_e^k\right) \right\} + S_N \sum_{e \in E'} \bar{f}_e \cdot \left(t_e - \bar{t}_e\right) + \frac{1}{2} \sum_{e \in E'} \left(t_e - \bar{t}_e\right)^2 \right\} \to \min_{t_e \ge \bar{t}_e,\, e \in E'}.$$

Stopping rule

$$(0\le)\,\Upsilon\left(\bar{t}^{\,N}\right)+\Psi\left(\bar{f}^{\,N}\right)\le\varepsilon. \qquad (*)$$

Theorem 3. *Let*

$$\tilde{M}_N^2=\left(\frac{1}{N+1}\sum_{k=0}^{N}M_k^{-2}\right)^{-1},\; R_N^2:=\frac{1}{2}\sum_{e\in E\setminus E'}\left(\tau_e\left(\bar{f}_e^{\,N}\right)-\bar{t}_e\right)^2+\frac{1}{2}\sum_{e\in E'}\left(\tilde{t}_e^{\,N}-\bar{t}_e\right)^2,$$

$$\left\{\tilde{t}_e^{\,N}\right\}_{e\in E'}=\arg\min_{\{t_e\}_{e\in E'}\ge 0}\left\{\underbrace{-\sum_{w\in W}d_wT_w\left(\left\{\tau_e\left(\bar{f}_e^{\,N}\right)\right\}_{e\in E\setminus E'},\{t_e\}_{e\in E'}\right)}_{F\left(\left\{\tau_e\left(\bar{f}_e^{\,N}\right)\right\}_{e\in E\setminus E'},\{t_e\}_{e\in E'}\right)}+\sum_{e\in E'}\bar{f}_e^{\,N}\cdot\left(t_e-\bar{t}_e\right)\right\}.$$

For arbitrary
$$N\ge\frac{2\tilde{M}_N^2R_N^2}{\varepsilon^2},$$

(*) is true and therefore

$$0\le\Upsilon\left(\bar{t}^{\,N}\right)-\Upsilon_*\le\varepsilon,\; 0\le\Psi\left(\bar{f}^{\,N}\right)-\Psi_*\le\varepsilon.$$

Another approach

$$\boxed{f_e^k \in -\partial_e F\left(t^k\right), \quad \bar{f}_e^N = \frac{1}{S_N}\sum_{k=0}^{N}\gamma_k f_e^k,\ e \in E,} \quad \tilde{R}^2 = \frac{1}{2}\sum_{e\in E'}\left(t_e^* - t_e^0\right)^2 = \frac{1}{2}\sum_{e\in E'}\left(t_e^* - \bar{t}_e\right)^2.$$

Theorem 4. *Let* $\tilde{R}_N^2 := \frac{1}{2}\sum_{e\in E\setminus E'}\left(\tau_e\left(\bar{f}_e^N\right) - \bar{t}_e\right)^2 + 5\tilde{R}^2$. *For arbitrary* $N \ge \frac{4\tilde{M}_N^2 \tilde{R}_N^2}{\varepsilon^2}$ *the following inequalities are satisfied*

$$\left|\Upsilon\left(\bar{t}^N\right) - \Upsilon_*\right| \le \varepsilon, \ \left|\Psi\left(\bar{f}^N\right) - \Psi_*\right| \le \varepsilon.$$

Moreover (stopping rule)

$$\sqrt{\sum_{e\in E'}\left(\left(\bar{f}_e^N - \bar{f}_e\right)_+\right)^2} \le \tilde{\varepsilon}, \tilde{\varepsilon} = \varepsilon / \tilde{R},$$

$$\Psi\left(\bar{f}^N\right) - \Psi_* \le \Upsilon\left(\bar{t}^N\right) + \Psi\left(\bar{f}^N\right) \le \varepsilon.$$

In arXiv:1701.02473 one can find how to solve the same problem with USTM.

Primal-dual method for Truss Topology Design (Nesterov–Shpirko)

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

We'd like to find such $\bar{x}^N$ that (see Lecture 3)

$$f\left(\bar{x}^N\right) - f_* \le \varepsilon_f = \frac{M_f}{M_g}\varepsilon_g,\ g\left(\bar{x}^N\right) \le \varepsilon_g,$$

$$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h_f \partial f\left(x^k\right)\right),\ \text{ if } g\left(x^k\right) \le \varepsilon_g,$$

$$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h_g \partial g\left(x^k\right)\right),\ \text{ if } g\left(x^k\right) > \varepsilon_g,$$

where $h_g = \varepsilon_g \big/ M_g^2,\ h_f = \varepsilon_g \big/ \left(M_f M_g\right),\ k = 1,\dots,N$. Let I be the set of such k that $g\left(x^k\right) \le \varepsilon_g,\ [N] = \{1,\dots,N\},\ J = [N]\setminus I,\ N_I = |I|,\ N_J = |J|,\ \bar{x}^N = \frac{1}{N_I}\sum_{k\in I} x^k$.

Let $g(x) = \max_{l=1,\ldots,m} g_l(x)$. Build a dual problem

$$\varphi(\lambda) = \min_{x \in Q} \left\{ f(x) + \sum_{l=1}^{m} \lambda_l g_l(x) \right\} \to \max_{\lambda \ge 0}.$$

Due to weak duality (see Lecture 1)

$$0 \le f(x) - \varphi(\lambda) \overset{def}{=} \Delta(x, \lambda),\ x \in Q,\ g(x) \le 0,\ \lambda \ge 0.$$

We assume that Slater's condition is true (Lect. 1): $\exists\ \tilde{x} \in Q :\ g(\tilde{x}) < 0$. Then

$$f_* = f(x_*) = \varphi(\lambda_*) = \varphi_*.$$

In this case the quality of approximate solution (x^N, λ^N) can be estimated by duality gap $\Delta(x^N, \lambda^N)$. The smaller is gap the better is solution.

Let

$$g\left(x^k\right)=g_{l(k)}\left(x^k\right),\ \partial g\left(x^k\right)=\partial g_{l(k)}\left(x^k\right),\ k\in J\,.$$

$$\boxed{\lambda_l^N=\frac{1}{h_f N_I}\sum_{k\in J}h_g I\left[l\left(k\right)=l\right]},\ I\left[\text{predicat}\right]=\begin{cases}1, & \text{predicat}=true\\ 0, & \text{predicat}=false\end{cases}.$$

Theorem 5. *Let* $\left\|\partial f\left(x\right)\right\|_*\le M_f$, $\left\|\partial g\left(x\right)\right\|_*\le M_g$ for all $x\in Q$.

Then for arbitrary

$$N\ge\frac{2M_g^2\bar{R}^2}{\varepsilon_g^2}+1.\ //\ \bar{R}^2=\max_{x,y\in Q}V\left(y,x\right)$$

the following inequalities are satisfied

$$N_I\ge 1\ and\ \Delta\left(\bar{x}^N,\bar{\lambda}^N\right)\le\varepsilon_f,\ g\left(\bar{x}^N\right)\le\varepsilon_g.$$

To be continued…