# Accelerating First Order Methods for Large-Scale Well-Structured Convex Optimization

Arkadi Nemirovski
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

*Joint research*
*with*
*Anatoli Juditsky*[†] *and Fatma Kilinc Karzan* [‡]
[†]: Joseph Fourier University, Grenoble, France; [‡]: ISyE, Georgia Tech

IPAM Workshop
"Numerical Methods for Continuous Optimization"
October 11-15, 2010

## Overview

- Goals
- Background:
  - Nesterov's strategy
  - Basic Mirror Prox algorithm
- Accelerating Mirror Prox:
  - Splitting
  - Utilizing strong concavity
  - Randomization

♣ **Problem:** Convex minimization problem

$$\mathrm{Opt}(P) = \min_{x\in X} f(x) \qquad \text{(P)}$$

• $X \subset \mathbf{R}^n$: convex compact • $f : x \to \mathbf{R}$: convex Lipschitz continuous

♣ **Goal:** We want to solve *nonsmooth large-scale* problems which, because of their sizes, are *beyond the "practical grasp" of polynomial time algorithms*

$\Rightarrow$*Focus on computationally cheap First Order methods with (nearly) dimension-independent rate of convergence,* meaning that for every $\epsilon > 0$, the number of First Order iterations needed to compute an $\epsilon$-solution $x_\epsilon \in X$:

$$f(x_\epsilon) - \mathrm{Opt}(P) \le \epsilon[\max_X f - \min_X f]$$

is bounded by $C \cdot M(\epsilon)$, where

- $M(\epsilon)$ is a *universal* (i.e., problem-independent) function
- $C$ is either an absolute constant, or a *universal* function of $n = \dim X$ with *slow* (e.g., logarithmic) growth.

$$\mathrm{Opt}(P) = \min_{x \in X} f(x) \qquad \text{(P)}$$

• $X \subset \mathbf{R}^n$: convex compact • $f : x \to \mathbf{R}$: convex Lipschitz continuous

**1.** Utilizing problem's structure, we represent $f$ as

$$f(x) = \max_{y \in Y} \phi(x,y)$$

• $Y \subset \mathbf{R}^m$: convex compact

• $\phi(x,y)$: convex in $x \in X$, concave in $y \in Y$ and *smooth*

$\Rightarrow$(P) becomes the convex-concave saddle point problem:

$$\mathrm{Opt}(P) = \min_{x \in X} \max_{y \in Y} \phi(x,y) \qquad \text{(SP)}$$

$$\Leftrightarrow \left\{ \begin{array}{ll} \mathrm{Opt}(P) = \min\limits_{x \in X} \left[ f(x) = \max\limits_{y \in Y} \phi(x,y) \right] & \text{(P)} \\ \mathrm{Opt}(D) = \max\limits_{y \in Y} \left[ \underline{f}(y) = \min\limits_{x \in X} \phi(x,y) \right] & \text{(D)} \end{array} \right.$$

$$\mathrm{Opt}(P) = \mathrm{Opt}(D)$$

$$\mathrm{Opt}(P) = \min_{x\in X} f(x) \;\Leftrightarrow\; \mathrm{Opt}(P) = \min_{x\in X}\max_{y\in Y}\phi(x,y)$$

**2.** (SP) is solved by a Saddle Point First Order method *utilizing smoothness of* $\phi$.
$\Rightarrow$*after* $t = 1, 2, \ldots$ *steps of the method, approximate solution* $(x^t, y^t) \in X \times Y$ *is built with*

$$f(x^t) - \mathrm{Opt}(P) \leq \varepsilon_{\rm sad}(x^t, y^t) := f(x^t) - \underline{f}(y^t) \leq O(1/t). \quad (!)$$

♣ **Note:** *When* $X, Y$ *are of "favorable geometry" and* $\phi$ *is "simple"* (which is the case in numerous applications),

- *Efficiency estimate (!) is "nearly dimension-independent:"*

$$\varepsilon_{\rm sad}(x^t, y^t) \leq C(\dim[X \times Y])\frac{\mathrm{Var}_X(f)}{t}, \;\; \mathrm{Var}_X(f) = \max_X f - \min_X f$$

- $C(n)$: grows with $n$ *at most logarithmically*

- *The method is "computationally cheap:" a step requires* $O(1)$ *computations of* $\nabla\phi$ *plus computational overhead of* $O(n)$ *("scalar case") or* $O(n^{3/2})$ *("matrix case") arithmetic operations.*

$$f(x^t) - \mathrm{Opt}(P) \leq O(1/t) \qquad \text{(!)}$$

♣ When solving *nonsmooth large-scale* problems, even "ideally structured" ones, by *First Order* methods, convergence rate $O(1/t)$ seems to be *unimprovable*. This is so already when solving Least Squares problems

$$\mathrm{Opt}(P) = \min_{x \in X} \left[ f(x) := \|Ax - b\|_2 \right], \, X = \{x \in \mathbf{R}^n : \|x\|_2 \leq R\}$$
$$\Leftrightarrow \mathrm{Opt}(P) = \min_{\|x\|_2 \leq R} \max_{\|y\|_2 \leq 1} y^T(Ax - b)$$

♣ **Fact** [Nem.'91]**:** *Given $t$ and $n > O(1)t$, for every method which generates $x^t$ after $t$ sequential calls to Multiplication oracle capable to multiply vectors, one at a time, by $A$ and $A^T$, there exists an $n$-dimensional Least Squares problem such that $\mathrm{Opt}(P) = 0$ and*

$$f(x^t) - \mathrm{Opt}(P) \geq O(1)\mathrm{Var}_X(f)/t.$$

- **Minimizing the maximum of smooth convex functions:**

$$\min_{x\in X}\max_{1\le i\le m} f_i(x)$$
$$\Leftrightarrow \min_{x\in X}\max_{y\in Y}\sum_i y_i f_i(x),\ Y=\{y\ge 0, \sum_i y_i = 1\}$$

- **Minimizing maximal eigenvalue:**

$$\min_{x\in X}\lambda_{\max}(\sum_i x_i A^i)$$
$$\Leftrightarrow \min_{x\in X}\max_{y\in Y}\mathrm{Tr}(y[\sum_i x_i A^i]),\ Y=\{y\succeq 0, \mathrm{Tr}(y)=1\}$$

- $\ell_1$ **minimization** $\min_{\xi}\{\|\xi\|_1 : \|A\xi - b\|_p \le \delta\}$ — *the* main tool in sparsity-oriented Signal Processing — reduces to a small series of parametric bilinear saddle point problems

$$\min_x\{\|Ax-\rho b\|_p : \|x\|_1\le 1\}$$
$$\Leftrightarrow \min_{x:\|x\|_1\le 1}\ \max_{y:\|y\|_q\le 1} y^T(Ax-\rho b),\ 1/p+1/q=1$$

- **Nuclear norm minimization**

$$\min_{X\in\mathbf{R}^{m\times n}} \{\|\sigma(X)\|_1 : \|\mathcal{A}(X)-b\|_p \le \epsilon\}$$

$$\left[\begin{array}{l} \bullet\ \|\mathcal{A}(\cdot) : \mathbf{R}^{m\times n} \to \mathbf{R}^N\text{: linear mapping} \\ \bullet\ \sigma(X)\text{: vector of singular values of } X \end{array}\right]$$

— the main tool of low rank approximation — reduces to small series of parametric bilinear saddle point problems

$$\min_{X\in\mathbf{R}^{m\times n}} \{\|\mathcal{A}(X)-\rho b\|_p : \|\sigma(X)\|_1 \le 1\}$$

$$\Leftrightarrow \min_{X:\|\sigma(X)\|_1\le 1}\ \max_{y:\|y\|_q\le 1} y^T(\mathcal{A}(X)-\rho b),\ \ 1/p+1/q=1$$

- **Uniform low-dimensional approximation** *"Given $N$ unit vectors $a_i \in \mathbf{R}^n$ and $k$, find subspace $L$, $\dim L = k$, minimizing $\max_i \mathrm{dist}_{\|\cdot\|_2}(a_i, L)$"* after SDP relaxation reduces to the bilinear saddle point problem

$$\min_{y\ge 0:\sum_i y_i=1}\ \max_{P:0\preceq P\preceq I,\mathrm{Tr}(P)\le k} \sum_i y_i a_i^T P a_i$$

$$\min_{x\in X}\max_{y\in Y}\phi(x,y) \qquad \text{(SP)}$$

- $X\subset E_x, Y\subset E_y$ : convex compacts in Euclidean spaces
- $\phi$ : convex-concave Lipschitz continuous

## MP Setup

♣ We fix:

• a norm $\|\cdot\|$ on the space $E=E_x\times E_y\supset Z:=X\times Y$

• a *distance-generating function* (d.-g.f.) $\omega(z):Z\to\mathbf{R}$ – a continuous convex function such that

— the subdifferential $\partial\omega(\cdot)$ admits a selection $\omega'(\cdot)$ continuous on $Z^o=\{z\in Z:\partial\omega(z)\neq\emptyset\}$

— $\omega(\cdot)$ is strongly convex modulus 1 w.r.t. $\|\cdot\|$:

$$\langle\omega'(z)-\omega'(z'),z-z'\rangle\geq\|z-z'\|^2\ \forall z,z'\in Z^o$$

♣ We introduce:

• $\omega$-center of $Z$: $z_\omega:=\mathrm{argmin}_Z\,\omega(\cdot)$

• Bregman distance: $V_z(u):=\omega(u)-\omega(z)-\langle\omega'(z),u-z\rangle$ $[z\in Z^o]$

• "$\omega$-size of $Z$": $\Omega:=\max_{u\in Z}V_{z_\omega}(u)$

• Prox-mapping: $\mathrm{Prox}_z(\xi)=\mathrm{argmin}_{u\in Z}\left[\langle\xi,u\rangle+V_z(u)\right]$ $[z\in Z^o,\xi\in E]$

$$\min_{x\in X}\max_{y\in Y}\phi(x,y) \qquad \text{(SP)}$$
$$F(x,y)=[F_x(x,y);F_y(x,y)]:Z=X\times Y\to E=E_x\times E_y:$$
$$F_x(x,y)\in\partial_x\phi(x,y),\ F_y(x,y)\in\partial_y[-\phi(x,y)]$$

## ♣ Basic MP algorithm:

$$\boxed{\begin{array}{rcl} z_1 &=& z_\omega := \mathrm{argmin}_Z\,\omega(\cdot) \\ \hline z_t \Rightarrow w_t &=& \mathrm{Prox}_{z_t}(\gamma_t F(z_t)) \quad [\gamma_t>0:\ \text{stepsizes}] \\ \Rightarrow z_{t+1} &=& \mathrm{Prox}_{z_t}(\gamma_t F(w_t)) \\ \hline z^t &=& (x^t,y^t) := \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \sum_{\tau=1}^t \gamma_\tau w_\tau \end{array}}$$

## Illustration: Euclidean setup

- $\|\cdot\|=\|\cdot\|_2$, $\omega(z)=\frac{1}{2}z^Tz$

$\Rightarrow V_z(u)=\frac{1}{2}\|u-z\|_2^2$, $\Omega=O(1)\max\limits_{u,v\in Z}\|u-v\|_2^2$, $\mathrm{Prox}_z(\xi)=\mathrm{Proj}_Z(z-\xi)$

$$\Rightarrow \boxed{\begin{array}{rcl} z_1 &\in& Z \\ \hline w_t &=& \mathrm{Proj}_Z(z_t-\gamma_t F(z_t)) \\ z_{t+1} &=& \mathrm{Proj}_Z(z_t-\gamma_t F(w_t)) \\ \hline z^t &=& \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \sum_{\tau=1}^t \gamma_\tau w_\tau \end{array}}$$

$$\min_{x\in X}\max_{y\in Y}\phi(x,y) \qquad \text{(SP)}$$

$$F(x,y)=[F_x(x,y);F_y(x,y)]:Z=X\times Y\to E=E_x\times E_y:$$
$$F_x(x,y)\in\partial_x\phi(x,y),\ F_y(x,y)\in\partial_y[-\phi(x,y)]$$

♣ **Theorem** [Nem.'04]: *Let $F$ be Lipschitz continuous:*

$$\|F(z)-F(z')\|_*\le L\|z-z'\|\ \forall z,z'\in Z,$$

*($\|\cdot\|_*$ is the conjugate of $\|\cdot\|$) and let $\gamma_\tau\ge L^{-1}$ be such that*

$$\gamma_\tau\langle F(w_\tau),w_\tau-z_{\tau+1}\rangle\le F_{z_\tau}(z_{\tau+1}),$$

*which definitely is the case when $\gamma_\tau\equiv L^{-1}$. Then*

$$\forall t\ge 1:\varepsilon_{\rm sad}(z^t)\le\left[\sum_{\tau=1}^t\gamma_\tau\right]^{-1}\Omega\le\Omega L/t$$

$$\min_{x\in X}\max_{y\in Y}\phi(x,y) \qquad \text{(SP)}$$

♣ Let $Z = X \times Y$ be a *subset* of the direct product $Z^+$ of $p+q$ *standard blocks*: $Z := X \times Y \subset Z^+ = Z^1 \times \ldots \times Z^{p+q}$

- $Z^i = \{\|z_i\|_2 \le 1\} \subset E_i = \mathbf{R}^{n_i}$, $1 \le i \le p$: *ball* blocks
- $Z^i = \mathcal{S}_i \subset E_i = \mathbf{S}^{\nu^i}$, $p+1 \le i \le p+q$: *spectahedron* blocks

$\mathbf{S}^{\nu^i}$: space of symmetric matrices of block-diagonal structure $\nu^i$ with the Frobenius inner product

$\mathcal{S}_i$: the set of all unit trace $\succeq 0$-matrices from $\mathbf{S}^{\nu^i}$

- $X$ and $Y$ are subsets of products of *complementary* groups of $Z^i$'s

♣ **Note:**

- The simplex $\Delta_n = \{x \in \mathbf{R}^n_+ : \sum_i x_i = 1\}$ is a spectahedron;
- $\ell_1$/nuclear norm balls (as in $\ell_1$/nuclear norm minimization) can be expressed via spectahedrons:

$$u \in \mathbf{R}^n, \|u\|_1 \le 1 \quad \Leftrightarrow \quad \exists [v,w] \in \Delta_{2n} : u = v - w$$

$$U \in \mathbf{R}^{p\times q}, \|U\|_* \le 1 \quad \Leftrightarrow \quad \exists V, W : \left[\begin{array}{c|c} V & \frac{1}{2}U \\ \hline \frac{1}{2}U^T & W \end{array}\right] \in \mathcal{S}$$

$$\min_{x\in X} \max_{y\in Y} \phi(x,y) \qquad \text{(SP)}$$
$$X\times Y := Z \subset Z^+ = Z^1 \times \ldots \times Z^{p+q}$$

♣ We associate with blocks $Z^i$ “partial MP setup data:”

| Block | Norm on the embedding space | d.-g.f. | $\omega_i$-size of $Z^i$ |
|---|---|---|---|
| ball $Z^i \subset \mathbf{R}^{n_i}$ | $\\|z_i\\|_{(i)} \equiv \\|z_i\\|_2$ | $\frac{1}{2}z_i^T z_i$ | $\Omega_i = \frac{1}{2}$ |
| spectahedron $Z^i \subset \mathbf{S}^{\nu^i}$ | $\\|z_i\\|_{(i)} \equiv \\|\lambda(z_i)\\|_1$ | $\sum_\ell \lambda_\ell(z_i)\ln\lambda_\ell(z_i)$ | $\Omega_i = \ln(\|\nu^i\|)$ |

$[\lambda_\ell(z_i)$ : eigenvalues of $z_i \in \mathbf{S}^{\nu^i}]$

♣ Assuming $\nabla\phi$ Lipschitz continuous, we find $L_{ij} = L_{ji}$ satisfying

$$\|\nabla_{z_i}\phi(u) - \nabla_{z_i}\phi(v)\|_{(i,*)} \le \sum_j L_{ij}\|u_j - v_j\|_{(j)}$$

♣ *Partial setup data induce MP setup for (SP) yielding the efficiency estimate*

$$\boxed{\forall t : \varepsilon_{\rm sad}(z^t) \le \mathcal{L}/t, \ \ \mathcal{L} = \sum_{i,j} L_{ij}\sqrt{\Omega_i\Omega_j}}$$

$$\min_{x\in X}\left[f(x)=\max_{y\in Y}\phi(x,y)\right] \qquad \text{(SP)}$$

• $Z := X\times Y \subset Z^+ = Z^1\times\ldots\times Z^{p+q}$

• $Z^1,\ldots,Z^p$: unit balls • $Z^{p+1},\ldots,Z^{p+q}$: spectahedrons

$\|\nabla_{z_i}\phi(u)-\nabla_{z_i}\phi(v)\|_{(i,*)}\le\sum_j L_{ij}\|u_j-v_j\|_{(j)}$

$$\Rightarrow\quad \boxed{\begin{array}{c}\varepsilon_{\rm sad}(z^t)\le\mathcal{L}/t,\\ \mathcal{L}=\sum_{i,j}L_{ij}\sqrt{\Omega_i\Omega_j}\le\ln(\dim Z)(p+q)^2\max_{i,j}L_{ij}\end{array}} \qquad (!)$$

♣ In good cases, $p+q=O(1)$, $\ln(\dim Z)\le O(1)\ln(\dim X)$ and $\max_{i,j}L_{ij}\le O(1)[\max_X f-\min_X f]$

$\Rightarrow$(!) *becomes nearly dimension-independent $O(1/t)$ efficiency estimate*

$$f(x^t)-\min_X f\le O(1)\ln(\dim X)\mathrm{Var}_X(f)/t$$

♣ *If $Z$ is cut off $Z^+$ by $O(1)$ linear inequalities, the effort per iteration reduces to $O(1)$ computations of $\nabla\phi$ and eigenvalue decomposition of $O(1)$ matrices from $\mathbf{S}^{\nu^i}$, $p+1\le i\le p+q$.*

$$\boxed{\begin{array}{c}\mathrm{Opt}(P)=\min_{\xi\in\Xi}\left[f(\xi)=\|A\xi-b\|_p\right],\ \Xi=\{\xi:\|\xi\|_\pi\leq R\}\\ \bullet\ A\text{: } m\times n\ \bullet\ p\text{: 2 or } \infty\ \bullet\ \pi\text{: 1 or 2}\end{array}}$$

$$\Updownarrow$$

$$\boxed{\mathrm{Opt}(P)=\min_{\|x\|_\pi\leq 1}\max_{\|y\|_{p_*}\leq 1} y^T(RAx-b),\ p_*=p/(p-1)}$$

♣ *Setting*

$$\|A\|_{\pi,p}=\max_{\|x\|_\pi\leq 1}\|Ax\|_p=\left\{\begin{array}{l}\max_{1\leq j\leq n}\|\mathrm{Column}_j(A)\|_p,\ \pi=1\\ \|\sigma(A)\|_\infty,\ \pi=p=2\\ \max_{1\leq i\leq m}\|\mathrm{Row}_i(A)\|_2,\ \pi=2,p=\infty\end{array}\right.$$

*the efficiency estimate of MP reads*

$$f(x^t)-\mathrm{Opt}(P)\leq O(1)[\ln(n)]^{\frac{1}{\pi}-\frac{1}{2}}[\ln(m)]^{\frac{1}{2}-\frac{1}{p}}\|A\|_{\pi,p}/t$$

♣ *When problem is "nontrivial:"* $\mathrm{Opt}(P)\leq\frac{1}{2}\|b\|_p$, *this implies*

$$f(x^t)-\mathrm{Opt}(P)\leq O(1)[\ln(n)]^{\frac{1}{\pi}-\frac{1}{2}}[\ln(m)]^{\frac{1}{2}-\frac{1}{p}}\mathrm{Var}_\Xi(f)/t$$

**Note:** When $\pi=1$, the results remain intact when passing from $\Xi=\{\xi\in\mathbf{R}^n:\|\xi\|_1\leq R\}$ to $\Xi=\{\xi\in\mathbf{R}^{n\times n}:\|\sigma(\xi)\|_1\leq R\}$.

♣ **Fact** [Nesterov'07,Beck&Teboulle'08,...]: *If the objective* $f(x)$ *in a convex problem* $\min_{x \in X} f(x)$ *is given as* $f(x) = g(x) + h(x)$, *where* $g$, $h$ *are convex, and*

— $g(\cdot)$ *is smooth,*

— $h(\cdot)$ *is perhaps nonsmooth, but "easy to handle,"*

*then* $f$ *can be minimized at the rate* $O(1/t^2)$ *— "as if" there were no nonsmooth component.*

♣ This fact admits saddle point extension.

## Situation

♣ **Problem of interest:**

$$\min_{x\in X}\max_{y\in Y}\phi(x,y) \quad [\Rightarrow \Phi(z)=\partial_x\phi(z)\times\partial_y[-\phi(z)]]$$

- $X\subset E_x, Y\subset E_y$: convex compacts in Euclidean spaces
- $\phi$: convex-concave continuous
- $E=E_x\times E_y, Z=X\times Y$: equipped with norm $\|\cdot\|$ and d.-g.f. $\omega(\cdot)$

♣ **Splitting Assumption:**

$$\Phi(z)\supset G(z)+\mathcal{H}(z)$$

• $G(\cdot):Z\to E$: single-valued Lipschitz: $\|G(z)-G(z')\|_*\leq L\|z-z'\|$
• $\mathcal{H}(z)$: monotone convex valued with closed graph and "easy to handle:" *Given $\alpha>0$ and $\xi$, we can easily find a strong solution to the variational inequality given by $Z$ and the monotone operator $\mathcal{H}(\cdot)+\alpha\omega'(\cdot)+\xi$, that is, find $\bar{z}\in Z$ and $\zeta\in\mathcal{H}(\bar{z})$ such that*

$$\langle\zeta+\alpha\omega'(\bar{z})+\xi, z-\bar{z}\rangle\geq 0\ \forall z\in Z$$

$$\min_{x\in X}\max_{y\in Y}\phi(x,y) \;\Rightarrow\; \Phi(z)=\partial_x\phi(z)\times\partial_y[-\phi(z)]$$
$$\Phi(z)\supset G(z)+\mathcal{H}(z)$$

- $\|G(z)-G(z')\|_* \le L\|z-z'\|$
- $\mathcal{H}$: monotone and easy to handle

## Modified MP algorithm:

$$\boxed{\begin{array}{rcl} z_1 & = & z_\omega := \mathrm{argmin}_Z\,\omega \\ \hline z_t & \mapsto & w_t\in Z,\ \zeta_t\in\mathcal{H}(w_t):\\ \multicolumn{3}{c}{\langle\omega'(w_t)+L^{-1}[G(z_t)+\zeta_t]-\omega'(z_t),z-w_t\rangle\ge 0\ \forall z\in Z}\\ z_{t+1} & = & \mathrm{Prox}_{z_t}(L^{-1}[G(w_t)+\zeta_t])\\ & := & \mathrm{argmin}_{z\in Z}\left[\omega(z)+\langle L^{-1}[G(w_t)+\zeta_t]-\omega'(z_t),z\rangle\right]\\ \hline z^t & = & t^{-1}\sum_{\tau=1}^t w_\tau \end{array}}$$

## Efficiency estimate:

$$\varepsilon_{\mathrm{sad}}(z^t)\le \Omega L/t,$$
$$\Omega=\max_{z\in Z}[\omega(z)-\omega(z_\omega)-\langle\omega'(z_\omega),z-z_\omega\rangle]\text{: }\omega\text{-size of }Z.$$

♣ **Dantzig selector** recovery in Compressed Sensing:

$$\min_\xi\{\|\xi\|_1 : \|A^T(A\xi - b)\|_\infty \leq \delta\} \quad [A \in \mathbf{R}^{m\times n}]$$

reduces to solving short series of problems of $\|\cdot\|_\infty$-fit:

$$\mathrm{P}(R) : \ \mathrm{Opt}(R) = \min_{\|\xi\|_1\leq R} \left[f(\xi) := \|A^T(A\xi - b)\|_\infty\right]$$

♣ Applying MP to the saddle point reformulation of P($R$)

$$\mathrm{SP}(R) : \min_{\|x\|_1\leq 1}\max_{\|y\|_1\leq 1} y^T[Hx - h] \ \ [H = RA^TA,\ h = A^Tb]$$

$$\Rightarrow F(x,y) = [Hy; h - Hx],$$

the efficiency estimate is $f(\xi^t) - \mathrm{Opt}(R) \leq \ln(n)[\max_i H_{ii}]/t$.

♣ In typical Compressed Sensing applications, the columns $A_j$ of $A$ have $\|A_j\|_2 \approx 1$ and are nearly orthogonal to each other:

$$\mu := \max_{i\neq j} |A_i^TA_j| \ll 1 \Rightarrow H_{ii} \approx R,\ i \neq j \Rightarrow |H_{ij}| \leq \mu R \ll R.$$

♣ Denoting by $C$, $D$ the off-diagonal and diagonal parts of $H$, we have

$$F(x,y) = G(x,y) + \mathcal{H}(x,y) \equiv [Cy; h - Cx] + [Dy; -Dx]$$

and $\mathcal{H}$ is easy to handle ($D$ is diagonal!)

*$\Rightarrow$Modified MP results in $f(\xi^t) - \mathrm{Opt}(R) \leq \ln(n)\mu R/t$, Basic MP — in $f(\xi^t) - \mathrm{Opt}(R) \leq \ln(n)R/t$.*

**Note:** Typically, $\mu$ is as small as $O(\sqrt{\ln(m)/m})$ !

## Situation:

♣ **Problem of interest:**

$\min_{x\in X}\max_{y\in Y}\phi(x,y) \quad [\Rightarrow \Phi(z)=\partial_x\phi(z)\times\partial_y[-\phi(z)]]$

- $X\subset E_x$: convex compact,
  $E_x, X$ equipped with $\|\cdot\|_x$ and d.-g.f. $\omega_x(x)$
- $Y\subset E_y=\mathbf{R}^m$: closed and convex,
  $E_y, Y$ equipped with $\|\cdot\|_2$ and d.-g.f. $\omega_y(y)=\frac{1}{2}y^Ty$ [for simplicity]
- $\phi$: continuous, convex in $x$ and *strongly concave in $y$*:
  $x\in X, y\pm h\in Y \Rightarrow 2\phi(x,y)-\phi(x,y+h)-\phi(x,y-h)\geq\kappa\|h\|_2^2$

♣ **Modified Splitting Assumption:**

$$\Phi(x,y)\supset G(x,y)+\mathcal{H}(x,y)$$

- $G(x,y)=[G_x(x,y);G_y(x,y)]:Z\to E=E_x\times E_y$:
single-valued Lipschitz with $G_x(x,y)$ depending solely on $y$
- $\mathcal{H}(x,y)$: monotone convex valued with closed graph and "easy to handle:" *Given* $\alpha>0$ $\beta>0$ *and* $\xi$, *we can easily find* $\bar z=(\bar x,\bar y)\in Z$ *and* $\zeta\in\mathcal{H}(\bar z)$ *such that*

$$\langle\zeta+\xi+[\alpha\omega_x'(\bar x);\beta\omega_y'(\bar y)],z-\bar z\rangle\geq 0\ \forall z\in Z$$

$$\min_{x\in X}\max_{y\in Y}\phi(x,y) \qquad \text{(SP)}$$

♣ **Fact:** *Under outlined assumptions, the efficiency estimate of properly implemented Modified MP can be improved from $O(1/t)$ to $O(1/t^2)$.*

♣ **Idea of acceleration:**

• The error bound of MP is proportional to the $\omega$-size of the domain $Z = X \times Y$

• When applying Modified MP to (SP), strong concavity of $\phi$ in $y$ results in a qualified convergence of $y^t$ to the $y$-component $y_*$ of a saddle point

$\Rightarrow$*Eventually the (upper bound) on the distance from $y^t$ to $y_*$ will be reduced by absolute constant factor.* When it happens, independence of $G_x$ of $x$ allows to rescale the problem and to proceed *as if the $\omega$-size of $Z$ were reduced by absolute constant factor*.

$$\min_{x\in X}\max_{y\in Y}\phi(x,y)$$

♣ $G(x,y)=[G_x(x,y);G_y(x,y)]$ is Lipschitz continuous with $G_x$ independent of $x$, whence for properly chosen $L_{xy}$, $L_{yy}$ and all $x,x'\in X, y,y'\in Y$:

$$\begin{array}{rcrcl}\|G_x(x,y)-G_x(x',y')\|_{x,*} & \leq & & & L_{xy}\|y-y'\|_2\\ \|G_y(x,y)-G_y(x',y')\|_{y,*} & \leq & L_{xy}\|x-x'\|_x & + & L_{yy}\|y-y'\|_2\end{array}$$

## Theorem [Ioud.&Nem.'10]

*Under the Strong Concavity and Modified Splitting assumptions, the Modified MP admits "staged" implementation as follows:*

• *Iterations (completely similar to those of the original algorithm) are split in **stages**, with the total of $M_k$ iterations at the first $k$ stages;*

• *Let $k_*=\min\{k\in\mathbf{Z}: k\geq 1, 2^{k/2}\geq kR\frac{L_{xy}\sqrt{2\Omega_x}}{L_{yy}+2\kappa}\}$ ($R$ is a priori upper bound on $2\|y_*\|_2$, $\Omega_x$ is the $\omega_x$-size of $X$), and let $z^k=(x^k,y^k)$ be the approximate solutions built after $k$ stages. Then*

$$\begin{array}{lcl}k<k_* & \Rightarrow & \varepsilon_{\rm sad}(z^k)\leq k2^{-k}R^2 \;\&\; M_k\leq O(1)\,[L_{yy}/\kappa+1]\,k\\ k\geq k_* & \Rightarrow & \varepsilon_{\rm sad}(z^k)\leq O(1)\Omega_x L_{xy}^2/[\kappa M_k^2]\end{array}$$

♣ **Problem of interest:**

$$\mathrm{Opt} = \min_{\xi\in\Xi}\left[f(\xi) = h(\xi) + \sum_{\ell=1}^{L} \mathrm{dist}^2_{\|\cdot\|_2}(P_\ell\xi - p_\ell, U_\ell + V_\ell)\right]$$

- $\Xi \subset E_\xi$: convex compact,
  $E_\xi, \Xi$ are equipped with $\|\cdot\|_\xi$ and d.-g.f. $\omega_\xi(\cdot)$
- $h(\xi) : \Xi \to \mathbf{R}$: convex, continuous, "easy to handle:" *given* $\alpha > 0, a$, *it is easy to find* $\mathrm{argmin}_\Xi\left[h(\xi) + a^T\xi + \alpha\omega_\xi(\xi)\right]$
- $U_\ell$: convex compacts with easily computable $\|\cdot\|_2$-projectors
- $V_\ell = B_\ell \cdot \{\lambda \in \mathbf{R}^{n_\ell}_+ : \sum_i \lambda_i = 1\}$: convex hulls of given finite sets

## Example: Lasso

♣ The simplest special case of the above setting is the Lasso problem

$$\mathrm{Opt} = \min_{\|\xi\|_1 \le R}\left[f(\xi) := \|\xi\|_1 + \|P\xi - p\|_2^2\right]$$

with added upper bound on $\|\xi\|_1$.

$$\mathrm{Opt} = \min_{\xi\in\Xi}\left[f(\xi) = h(\xi) + \sum_{\ell=1}^{L} \mathrm{dist}^2_{\|\cdot\|_2}\left(P_\ell\xi - p_\ell, U_\ell + V_\ell\right)\right]$$

$$\bullet\ V_\ell = B_\ell \cdot \{\lambda \in \mathbf{R}^{n_\ell}_+ : \sum_i \lambda_i = 1\}$$

♣ *With the outlined approach, the efficiency estimate is*

$$f(\xi^k) - \mathrm{Opt} \leq O(1)\frac{\Omega_\xi \sum_{\ell=1}^{L}\|P_\ell\|^2 + \sum_{\ell=1}^{L}\|B_\ell\|^2_{1,2}\ln(n_\ell+1)}{M_k^2},\ k \geq k_*,$$

$$\bullet\ \Omega_\xi\text{: } \omega_\xi\text{-size of } \Xi \quad \bullet\ \|P_\ell\| = \max\{\|P_\ell\xi\|_2 : \|\xi\|_\xi \leq 1\}$$

*where $k_*$ is logarithmic in the magnitude of the data.*

♣ *Building $\xi^k$ reduces to $O(1)M_k$ computations of:*

*— solutions to auxiliary problems* $\min_{\xi\in\Xi}\left[h(\xi) + a^T\xi + \alpha\omega_\xi(\xi)\right]$,

*— matrix-vector products involving $P_\ell$, $P_\ell^T$, $B_\ell$, $B_\ell^T$, $1 \leq \ell \leq L$,*

*— projections of given points onto $U_\ell$, $1 \leq \ell \leq L$.*

♣ **Example:** For Lasso, we get $f(\xi^k) - \mathrm{Opt} \leq O(1)\ln(\dim\xi)\frac{R^2\|P\|^2_{1,2}}{M_k^2}$.

♣ **Note:** In terms of its efficiency and application scope, the outlined acceleration is similar to the "excessive gap technique" [Nesterov'05].

♣ We have seen that many important convex programs reduce to bilinear saddle point problems

$$\min_{x\in X}\max_{y\in Y}\left[\phi(x,y)=\langle a,x\rangle+\langle b,y\rangle+\langle y,Ax\rangle\right]$$ (

$$\Rightarrow F(z=(x,y))=[a;-b]+\mathcal{A}z,\ \mathcal{A}=\left[\begin{array}{c|c} & A^* \\ \hline -A & \end{array}\right]=-\mathcal{A}^*$$

♣ When $X,Y$ are simple, the computational cost of an iteration of a First Order method (e.g., MP) is dominated by computing $O(1)$ matrix-vector products $X\ni x\mapsto Ax$, $Y\ni y\mapsto A^*y$.

• *Can we save on computing these products?*

♣ Computing matrix-vector product $u \mapsto Bu : \mathbf{R}^p \to \mathbf{R}^q$ is easy to randomize, e.g., as follows:

> *pick a sample* $\jmath \in \{1,...,p\}$ *from the probability distribution* $\mathrm{Prob}\{\jmath = j\} = |u_j|/\|u\|_1$, $j = 1,...,p$ *and return* $\zeta = \|u\|_1 \mathrm{sign}(u_\jmath)\mathrm{Column}_\jmath[B]$.

♣ **Note:**

- $\zeta$ is an unbiased random estimate of $Bu$: $\mathbf{E}\{\zeta\} = Bu$;
- We have $\|\zeta\| \leq \|u\|_1 \max_j \|\mathrm{Column}_j[B]\|$
  $\Rightarrow$ *"noisiness" of the estimate is controlled by* $\|u\|_1$
- When the columns of $B$ are readily available, *computing* $\zeta$ *is simple:* given $u$, it takes

— $O(p)$ a.o. to compute $\{\sum_{j=1}^k |u_j|\}_{k=1}^p$ (setup cost),
— $O(\ln(p))$ a.o. to get a sample $\jmath$ after the setup cost is paid, and
— $O(q)$ a.o. to convert $\jmath$ into $\zeta$,
the total effort being $O(1)(p+q)$ a.o. (vs. $O(1)pq$ a.o. required for precise computation of $Bu$ for a general-type $B$).

$$\min_{x\in X}\max_{y\in Y}\left[\phi(x,y)=\langle a,x\rangle+\langle b,y\rangle+\langle y,Ax\rangle\right] \qquad \text{(SP)}$$

## ♣ Situation:

- $X\subset E_x$: convex compact, $E_x, X$ are equipped with $\|\cdot\|_x$ and d.-g.f. $\omega_x(\cdot)$
- $Y\subset E_y$: convex compact, $E_y, Y$ are equipped with $\|\cdot\|_y$ and d.-g.f. $\omega_y(\cdot)$

$$\Rightarrow\ \left\{\begin{array}{l}\Omega_x,\Omega_y:\ \text{respective } \omega\text{-sizes of } X, Y\\ \|A\|_{x,y}:=\max_x\{\|Ax\|_{y,*}: \|x\|_x\leq 1\}\end{array}\right.$$

- $x\in X$ are associated with probability distributions $P_x$ on $X$ such that $\mathbf{E}_{\xi\sim P_x}\{\xi\}\equiv x$
- $y\in Y$ are associated with probability distributions $\Pi_y$ on $E_y$ such that $\mathbf{E}_{\eta\sim \Pi_y}\{\eta\}\equiv y$.

$$\Rightarrow\ \left\{\begin{array}{l}\xi_u=\frac{1}{k_x}\sum_{\ell=1}^{k_x}\xi^\ell,\ \xi^\ell\sim P_u\text{: i.i.d. } [u\in X]\\ \eta_v=\frac{1}{k_y}\sum_{\ell=1}^{k_y}\eta^\ell,\ \eta^\ell\sim \Pi_v\text{: i.i.d. } [v\in Y]\\ \sigma_x^2=\sup_{u\in X}\mathbf{E}\{\|A[\xi_u-u]\|_{y,*}^2\}\\ \sigma_y^2=\sup_{v\in Y}\mathbf{E}\{\|A^*[\eta_v-v]\|_{x,*}^2\}\end{array}\right.$$

$$\Rightarrow\ \left\{\ \omega(x,y)=\frac{1}{2\Omega_x}\omega_x(x)+\frac{1}{2\Omega_y}\omega_y(y),\ \Theta=2\left[\Omega_x\sigma_y^2+\Omega_y\sigma_x^2\right]\right.$$

$$\min_{x\in X}\max_{y\in Y}\left[\phi(x,y)=\langle a,x\rangle+\langle b,y\rangle+\langle y,Ax\rangle\right] \qquad \text{(SP)}$$

$$[F(x,y)=[F_x=a+A^*y;F_y=-b-Ax]]$$

$$\|\cdot\|_x,\omega_x(\cdot),\|\cdot\|_y,\omega_y(\cdot),\{P_u\}_{u\in X},\{\Pi_v\}_{v\in Y},k_x,k_y$$

$\Rightarrow$ .........

$\Rightarrow$ $\{\xi_x,\, x\in X\};\ \{\eta_y, y\in Y\};\ \omega(x,y): Z:=X\times Y\to\mathbf{R};\ \ \Omega_x,\Omega_y,\Theta$

## Randomized MP Algorithm

♣ With number $N$ of steps given, set $\gamma=\min\left[\frac{1}{2\|A\|_{x,y}\sqrt{3\Omega_x\Omega_y}},\frac{1}{\sqrt{3\Theta N}}\right]$ and execute:

$$\boxed{z_1 = \mathrm{argmin}_{z\in Z}\,\omega(z)}$$

For $t=1,2,...,N$:

$$\begin{array}{rrcl}
& z_t=(x_t,y_t) & \Rightarrow & \zeta_t=[\xi_{x_t},\xi_{y_t}]\Rightarrow F(\zeta_t)\\
\Rightarrow & w_t=(u_t,v_t) & = & \mathrm{Prox}_{z_t}(\gamma F(\zeta_t))\\
& & := & \mathrm{argmin}_{w\in Z}\left\{\omega(w)+\langle\gamma F(\zeta_t)-\omega'(z_t),w\rangle\right\}\\
& & \Rightarrow & \widehat{\zeta}_t=[\xi_{u_t};\eta_{v_t}]\Rightarrow F(\widehat{\zeta}_t)\\
\Rightarrow & z_{t+1} & = & \mathrm{Prox}_{z_t}(\gamma F(\widehat{\zeta}_t))
\end{array}$$

$$\boxed{z^N=(x^N,y^N)=\frac{1}{N}\sum_{t=1}^N\widehat{\zeta}_t\Rightarrow F(z^N)=\frac{1}{N}\sum_{t=1}^N F(\zeta_t).}$$

$$\mathrm{Opt} = \min_{x\in X}\left\{f(x) := \max_{y\in Y}\left[\langle a,x\rangle + \langle b,y\rangle + \langle y,Ax\rangle\right]\right\} \qquad \text{(SP)}$$

$\Rightarrow \ldots\ldots\ldots \Rightarrow \Omega_x, \Omega_y, \Theta$

**Theorem [Ioud.&Nem.'10]**

*For every $N$, the $N$-step Randomized MP algorithm ensures that $x^N \in X$ and*

$$\mathbf{E}\left\{f(x^N) - \mathrm{Opt}\right\} \le \max\left[\frac{2\sqrt{2\Theta}}{\sqrt{N}}, \frac{4\sqrt{3}\|A\|_{x,y}\sqrt{\Omega_x\Omega_y}}{N}\right].$$

*When $\Pi_y$ is supported on $Y$ for all $y \in Y$, then also $y^N \in Y$ and*

$$\mathbf{E}\left\{\varepsilon_{\mathrm{sad}}(z^N)\right\} \le \max\left[\frac{2\sqrt{3\Theta}}{\sqrt{N}}, \frac{4\sqrt{3}\|A\|_{x,y}\sqrt{\Omega_x\Omega_y}}{N}\right].$$

**Note:** The method produces both $z^N$ and $F(z^N)$, which allows for easy computation of $\varepsilon_{\mathrm{sad}}(z^N)$. This feature is instrumental when Randomized MP is used as "working horse" in processing, e.g., $\ell_1$ minimization problems

$$\min_x \left\{\|x\|_1 : \|Ax - b\|_p \le \delta\right\}$$

♣ $\ell_1$ minimization with uniform fit

$$\min_\xi \{\|\xi\|_1 : \|A\xi - b\|_\infty \leq \delta\} \quad [A : m \times n]$$

reduces to a small series of problems

$$\begin{array}{rcl} \mathrm{Opt} & = & \min_{\|x\|_1 \leq 1} \|Ax - \rho b\|_\infty \\ & = & \min\limits_{\|x\|_1 \leq 1} \max\limits_{\|y\|_1 \leq 1} y^T(Ax - \rho b) \end{array} \qquad (!)$$

## Corollary of Theorem:

*For every $N$, one can find random feasible solution $(x^N, y^N)$ to $(!)$, along with $Ax^N$, $A^T y^N$, in such a way that*

$$\mathrm{Prob}\left\{\varepsilon_{\mathrm{sad}}(x^N, y^N) \leq O(1)\frac{\ln(2mn)\max_{i,j}|A_{ij}|}{\sqrt{N}}\right\} > \frac{1}{2}$$

*in $N$ steps of Randomized MP, with effort per step dominated by extracting from $A$ $O(1)$ columns and rows, given their indices.*

$$\begin{array}{rcl} \mathrm{Opt} & = & \min_{\|x\|_1\le 1} \|Ax - \rho b\|_\infty \\ & = & \min\limits_{\|x\|_1\le 1} \max\limits_{\|y\|_1\le 1} y^T(Ax-\rho b) \end{array} \qquad (!)$$

♣ Let confidence level $1-\beta$, $\beta \ll 1$ and $\epsilon < \max_{i,j}|A_{ij}|$ be given. Applying Randomized MP, we with confidence $\geq 1-\beta$ find a feasible solution $(\bar{x},\bar{y})$ satisfying $\varepsilon_{\mathrm{sad}}(\bar{x},\bar{y}) \leq \epsilon$ in

$$O(1)\ln^2(2mn)\ln(1/\beta)(m+n)\left[\frac{\max_{i,j}|A_{ij}|}{\epsilon}\right]^2$$

arithmetic operations.

♣ When $A$ is general type dense $m \times n$ matrix, the best known complexity of finding $\epsilon$-solution to (!) by a deterministic algorithm is, for $\epsilon$ fixed and $m, n$ large,

$$O(1)\sqrt{\ln(2m)\ln(2n)}mn\left[\frac{\max_{i,j}|A_{ij}|}{\epsilon}\right]$$

arithmetic operations.

$\Rightarrow$*When the relative accuracy* $\epsilon/\max_{i,j}|A_{ij}|$ *is fixed and* $m, n$ *are large, the computational effort in the randomized algorithm is negligible as compared to the one in a deterministic method.*

$$\begin{array}{rcl} \mathrm{Opt} & = & \min_{\|x\|_1\le 1}\|Ax-\rho b\|_\infty \\ & = & \min\limits_{\|x\|_1\le 1}\max\limits_{\|y\|_1\le 1} y^T(Ax-\rho b) \end{array} \qquad (!)$$

♣ The efficiency estimate

$$O(1)\ln^2(2mn)\ln(1/\beta)(m+n)\left[\frac{\max_{i,j}|A_{ij}|}{\epsilon}\right]^2 \text{ a.o.}$$

says that *with $\epsilon, \beta$ fixed and $m, n$ large, the Randomized MP exhibits sublinear time behavior: $\epsilon$-solution is found reliably while looking through a negligible fraction of the data.*

**Note:** (!) is equivalent to a zero sum matrix game, and a such can be solved by the sublinear time randomized algorithm for matrix games [Grigoriadis&Khachiyan'95]. In hindsight, this "ad hoc" algorithm is close, although not identical, to Randomized MP as applied to (!).

♣ $\ell_1$ minimization with $\|\cdot\|_2$ fit

$$\min_\xi \{\|\xi\|_1 : \|A\xi - b\|_2 \leq \delta\} \quad [A : m \times n]$$

reduces to a small series of problems

$$\begin{array}{rcl} \mathrm{Opt} & = & \min_{\|x\|_1\leq 1} \|Ax - \rho b\|_2 \\ & = & \min\limits_{\|x\|_1\leq 1} \max\limits_{\|y\|_2\leq 1} y^T(Ax - \rho b) \end{array} \qquad (!)$$

## Corollary of Theorem:

*For every $N$, one can find random feasible solution $x^N$ to (!), along with $Ax^N$, such that*

$$\mathrm{Prob}\left\{\|Ax^N - \rho b\|_2 - \mathrm{Opt} \leq O(1)\frac{\sqrt{\ln(2n)}\|A\|_{1,2}\Gamma(A)}{\sqrt{N}}\right\} \geq \frac{1}{2},$$

$$\|A\|_{1,2} = \max_j \|\mathrm{Column}_j[A]\|_2, \ \Gamma(A) = \sqrt{m}\|A\|_{1,\infty}/\|A\|_{1,2}.$$

*in $N$ steps of Randomized MP, with effort per step dominated by extracting from $A$ $O(1)$ columns and rows, given their indices.*

$$\mathrm{Prob}\left\{\|Ax^N-\rho b\|_2-\mathrm{Opt}\leq O(1)\frac{\sqrt{\ln(2n)}\|A\|_{1,2}\Gamma(A)}{\sqrt{N}}\right\}\geq\frac{1}{2},$$
$$\|A\|_{1,2}=\max_j\|\mathrm{Column}_j[A]\|_2,\ \ \Gamma(A)=\sqrt{m}\|A\|_{1,\infty}/\|A\|_{1,2}.$$

**Remark:**

♣ $\Gamma(A)$ can be as large as $\sqrt{m}$.

**However:** *Randomized preprocessing*

$$[A,b]\Rightarrow[\widetilde{A},\widetilde{b}]=U\ \mathrm{Diag}\{\xi\}[A,b]$$

- $U$: *orthogonal easy-to-multiply matrix with* $|U_{ij}|\leq O(1)/\sqrt{m}$
- $\xi$: *random* $\sim\mathrm{Uniform}\left(\{-1,1\}^m\right)$

*results in equivalent problem and with confidence* $1-\beta$ *makes* $\Gamma$ *as small as* $O(1)\sqrt{\ln(mn/\beta)}$.

♣ *The cost of preprocessing is* $O(1)mn\ln(m)$ *a.o.*

- A. Beck, M. Teboulle, A Fast Iterative... – *SIAM J. Imag. Sci.* '08
- D. Goldfarb, K. Scheinberg, Fast First Order... Tech. rep. Dept. IEOR, Columbia Univ. '10
- M. Grigoriadis, L. Khachiyan, A Sublinear Time... – *OR Letters* **18** '95
- A. Juditsky, F. Kilinç Karzan, A. Nemirovski, $\ell_1$ Minimization... ('10), http://www.optimization-online.org
- A. Juditsky, A. Nemirovski, First Order... I,II: to appear in: S. Sra, S. Novozin, S.J. Wright, Eds., *Optimization for Machine Learning*, MIT
- A. Nemirovski, Information-Based... – *J. of Complexity* **8** '92
- A. Nemirovski, Prox-Method... – *SIAM J. Optim.* **15** '04
- Yu. Nesterov, A Method for Solving... – *Soviet Math. Dokl.* **27:2** '83
- Yu. Nesterov, Smooth Minimization... – *Math. Progr.* **103** '05
- Yu. Nesterov, Excessive Gap Technique... *SIAM J. Optim.* **16:1** '05
- Yu. Nesterov, Gradient Methods for Minimizing... CORE Discussion Paper '07/76