Saddle Point Algorithms 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; [‡]: Tepper Business School,
Carnegie Mellon University, Pittsburgh, USA

Institut Henri Poincaré
OPTIMIZATION, GAMES, AND DYNAMICS
November 28-29, 2011

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:** to solve *nonsmooth large-scale problems of sizes beyond the "practical grasp" of polynomial time algorithms*
⇒*Focus on computationally cheap First Order methods with (nearly) dimension-independent rate of convergence:*

- for every $\epsilon > 0$, an ϵ-solution $x_\epsilon \in X$:
$$f(x_\epsilon) - \mathrm{Opt}(P) \leq \epsilon[\max_X f - \min_X f]$$
is computed in at most $C \cdot M(\epsilon)$ First Order iterations, 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:*

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

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

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

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

2. (SP) is solved by a Saddle Point First Order method *utilizing smoothness of* ϕ.
$\Rightarrow$*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* ϕ *is "simple"* (which is the case in numerous applications),

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

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

- $C(n)$: 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.*

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

♣ 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

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

♣ **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*

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

• **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:**

$$\min_{x\in X}\lambda_{\max}(\textstyle\sum_i x_i A^i)$$

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

• **L_1/Nuclear norm minimization.** *The* main tool in sparsity oriented Signal Processing – the problem

$$\min_\xi\{\|\xi\|_1 : \|A(\xi)-b\|_p\le\delta\}$$

• $\xi\mapsto A(\xi)$: linear • $\|\cdot\|_1$: ℓ_1/nuclear norm of a vector/matrix

reduces to a small series of bilinear saddle point problems

$$\min_x\{\|A(x)-\rho b\|_p : \|x\|_1\le 1\}\Leftrightarrow \min_{\|x\|_1\le 1}\max_{\|y\|_{p/(p-1)}\le 1} y^T(A(x)-\rho b)$$

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

- $X \subset E_x, Y \subset E_y$: convex compacts in Euclidean spaces
- ϕ : 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:

- ω-center of Z: $z_\omega := \mathrm{argmin}_Z\, \omega(\cdot)$
- Bregman distance: $V_z(u) := \omega(u) - \omega(z) - \langle\omega'(z), u - z\rangle$ $[z \in Z^o]$
- Prox-mapping: $\mathrm{Prox}_z(\xi) = \mathrm{argmin}_{u\in Z}\,[\langle\xi, u\rangle + V_z(u)]$ $[z \in Z^o, \xi \in E]$
- "ω-size of Z": $\Omega := \max_{u\in Z} V_{z_\omega}(u)$

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

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

$$F_x(x,y)\in\partial_x\phi(x,y),\ F_y(x,y)\in\partial_y[-\phi(x,y)]$$

♣ Basic MP algorithm:

$$\boxed{z_1 = z_\omega := \mathrm{argmin}_Z\,\omega(\cdot)}$$

$$z_t \Rightarrow w_t = \mathrm{Prox}_{z_t}(\gamma_t F(z_t)) \quad [\gamma_t>0:\ \text{stepsizes}]$$

$$\Rightarrow z_{t+1} = \mathrm{Prox}_{z_t}(\gamma_t F(w_t))$$

$$z^t = (x^t,y^t) := \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1}\sum_{\tau=1}^t \gamma_\tau w_\tau$$

Illustration: Euclidean setup

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

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

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

Note: Up to averaging, this is *Extragradient method* due to G. Korpelevich '76.

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

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

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

  $\mathbf{S}^{\nu^i}$: space of symmetric matrices of block-diagonal structure ν^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;
- ℓ_1/nuclear norm balls (as in ℓ_1/nuclear norm minimization) can be expressed via spectahedrons:

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

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

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

♣ We associate with blocks Z^i "partial MP setup data:"

Block	Norm on the embedding space	d.-g.f.	ω_i-size of Z^i
ball $Z^i\subset \mathbf{R}^{n_i}$	$\lVert z_i\rVert_{(i)}\equiv\lVert z_i\rVert_2$	$\frac{1}{2}z_i^Tz_i$	$\Omega_i=\frac{1}{2}$
spectahedron $Z^i\subset \mathbf{S}^{\nu^i}$	$\lVert z_i\rVert_{(i)}\equiv\lVert\lambda(z_i)\rVert_1$	$\sum_\ell\lambda_\ell(z_i)\ln\lambda_\ell(z_i)$	$\Omega_i=\ln(\lvert\nu^i\rvert)$

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

♣ Assuming $\nabla\phi$ Lipschitz continuous, we find $L_{ij}=L_{ji}$ satisfying
$$\lVert\nabla_{z_i}\phi(u)-\nabla_{z_i}\phi(v)\rVert_{(i,*)}\leq\sum_j L_{ij}\lVert u_j-v_j\rVert_{(j)}$$

♣ *Partial setup data induce MP setup for (SP) yielding the efficiency estimate*
$$\boxed{\forall t:\ \varepsilon_{\rm sad}(z^t)\leq\mathcal{L}/t,\quad \mathcal{L}=\sum_{i,j}L_{ij}\sqrt{\Omega_i\Omega_j}}$$

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

- $Z:=X\times Y\subset Z^+=Z^1\times\ldots\times Z^{p+q}$
- $Z^1,\ldots,Z^p$: unit balls • $Z^{p+1},\ldots,Z^{p+q}$: spectahedrons

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

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

♣ In good cases, $p+q=O(1)$, $\ln(\dim Z)\leq O(1)\ln(\dim X)$ and $\max_{i,j}L_{ij}\leq O(1)[\max_X f-\min_X f]$
$\Rightarrow$(!) *becomes nearly dimension-independent $O(1/t)$ efficiency estimate*

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

♣ *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 $\mathbf{S}^{\nu^i}$, $p+1\leq i\leq p+q$.*

$$\mathrm{Opt}(P) = \min_{\xi\in\Xi}\left[f(\xi) = \|A\xi - b\|_p\right],\ \Xi = \{\xi : \|\xi\|_\pi \le R\}$$

• A: $m \times n$ • p: 2 or ∞ • π: 1 or 2

$$\Updownarrow$$

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

♣ *Setting*

$$\|A\|_{\pi,p} = \max_{\|x\|_\pi\le 1}\|Ax\|_p = \begin{cases} \max_{1\le j\le n}\|\mathrm{Column}_j(A)\|_p, & \pi = 1\\ \|\sigma(A)\|_\infty, & \pi = p = 2\\ \max_{1\le i\le m}\|\mathrm{Row}_i(A)\|_2, & \pi = 2, p = \infty\end{cases}$$

the efficiency estimate of MP reads

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

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

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

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

$$\widehat{x} \approx \underset{x}{\operatorname{argmin}} \{\|Ax - b\|_\infty : \|x\|_1 \leq 1\}$$

A: random $m \times n$ submatrix of $n \times n$ D.F.T. matrix
b: $\|Ax_* - b\|_\infty \leq \delta = 5.e\text{-}3$ with 16-sparse x_*, $\|x_*\|_1 = 1$

		Errors			CPU
$m \times n$	Method	$\|x_* - \widehat{x}\|_1$	$\|x_* - \widehat{x}\|_2$	$\|x_* - \widehat{x}\|_\infty$	sec
512×2048	DMP	0.0052	0.0018	0.0013	3.3
	IP	0.0391	0.0061	0.0021	321.6
1024×4096	DMP	0.0096	0.0028	0.0015	3.5
	IP	Out of space (2GB RAM)			
4096×16384	DMP	0.0057	0.0026	0.0024	46.4
	IP	not tested			

- Mirror Prox utilizing FFT
- IP: Commercial Interior Point LP solver `mosekopt`

♣ **Situation and Goal:** We observe 33% of randomly selected pixels in a 256×256 image X and want to recover the entire image.

♠ **Solution strategy:** Representing the image in a wavelet basis: $X = Ux$, the observation becomes $y = Ax$, where A is comprised of randomly selected rows of U.

Applying the ℓ_1 minimization, the recovered image is $\widehat{X} = U\widehat{x}$,

$$\widehat{x} = \operatorname*{Argmin}_x \{\|x\|_1 : Ax = b\}$$

Note: multiplication of a vector by A and A^T takes linear time

$\Rightarrow$situation is perfectly well suited for First Order methods

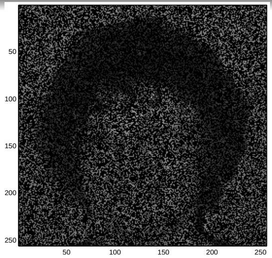
♠ Matrix A:

- sizes $21,789 \times 65,536$
- density 4% (5.3×10^7 nonzero entries)

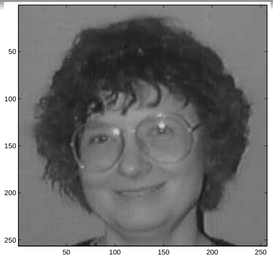
♠ Target accuracy: we seek for $\widetilde{x}$ such that $\|\widetilde{x}\|_1 \leq \|\widehat{x}\|_1$ and $\|A\widetilde{x} - b\|_2 \leq 0.0075\|b\|_2$

♠ CPU time: 1,460 sec (MATLAB, 2.13 GHz single-core Intel Pentium M processor, 2 GB RAM)

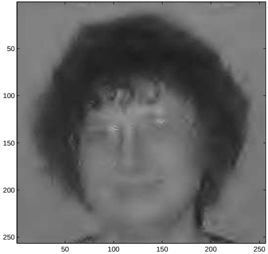
Example 2 (continued)



Observations

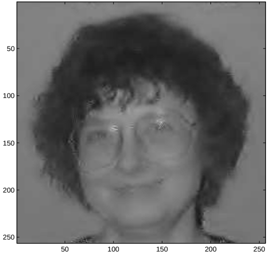


True image



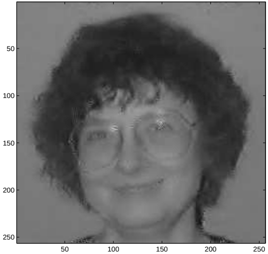
Steps: 328 CPU: $99''$

$\frac{\|Ax-b\|_2}{\|b\|_2} = 0.0647$



Steps: 947 CPU: $290''$

$\frac{\|Ax-b\|_2}{\|b\|_2} = 0.0271$



Steps: 4,746 CPU: $1460''$

$\frac{\|Ax-b\|_2}{\|b\|_2} = 0.0075$

♣ **Problem:** Given graph G with n nodes and m arcs, compute

$$\theta(G) = \min_{X \in \mathbf{S}^n} \{\lambda_{\max}(X + \mathbf{J}) : X_{ij} = 0 \text{ when } (i,j) \text{ is an arc}\}$$

within accuracy ϵ.

- **J**: all-ones matrix

♣ **Saddle point reformulation:**

$$\min_{X \in \mathcal{X}} \max_{Y \in \mathcal{Y}} \mathrm{Tr}(Y(X + \mathbf{J}))$$

$$\left[\begin{array}{lcl} \mathcal{X} & = & \{X \in \mathbf{S}^n : X_{ij} = 0 \text{ when } (i,j) \text{ is an arc}, |X_{ij}| \leq \bar{\theta}\} \\ \mathcal{Y} & = & \{Y \in \mathbf{S}^n : Y \succeq 0, \mathrm{Tr}(Y) = 1\} \\ \bar{\theta} & : & \text{a priori upper bound on } \theta(G) \end{array}\right]$$

♠ For ϵ fixed and n large, theoretical complexity of estimating $\theta(G)$ within accuracy ϵ is *by orders of magnitude smaller* than the cost of a *single* IP iteration.

# of arcs	# of nodes	# of steps, $\epsilon = 1$	CPU time, Mirror Prox	CPU time, IPM (estimate)
616	50	527	$2''$	0
2,459	100	738	$15''$	15 sec
4,918	200	1,003	$2'\ 30''$	$>$2 min
11,148	300	3,647	$32'\ 08''$	$>$23 min
20,006	400	2,067	$46'\ 35''$	$>$2 hours
62,230	500	1,867	$25'\ 21''$	$>$2.7 days
197,120	1024	1,762	$1^h\ 37'\ 40''$	$>$12.7 weeks

Computing Lovasz Capacity, performance 3 Gfl/sec.

♣ **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
- ϕ: 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 ξ, 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*

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

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

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

♣ **Theorem** [Ioud.&Nem.'11]: *Under Splitting Assumption, the MP algorithm can be modified to yield the efficiency estimate "as if" there were no $\mathcal{H}$-component:*

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

An iteration of the modified algorithm costs 2 computations of $G(\cdot)$, solving auxiliary problem as in Splitting Assumption, and computing 2 prox-mappings.

♣ **Dantzig selector** recovery in Compressed Sensing reduces to solving the problem

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

• In typical Compressed Sensing applications, the diagonal entries in A^TA are $O(1)$'s, while moduli of off-diagonal entries do not exceed $\mu \ll 1$ (usually, $\mu = O(1)\sqrt{\ln(n)/m}$).

⇒*In the saddle point reformulation of Dantzig selector problem, splitting induced by partitioning A^TA into its off-diagonal and diagonal parts accelerates the solution process by factor $1/\mu$.*

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 equipped with $\|\cdot\|_y$ and d.-g.f. $\omega_y(y)$
- ϕ: continuous, convex in x and
  strongly concave in y w.r.t. $\|\cdot\|_y$

♣ **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."

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

♣ **Fact** [Ioud.&Nem'11]: *Under outlined assumptions, the efficiency estimate of properly 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 ω-size of the domain $Z = X \times Y$

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

⇒*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 ω-size of Z were reduced by absolute constant factor*.

♣ **Problem of interest:**

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

(LASSO with added upper bound on $\|\xi\|_1$).

♣ **Result:** *With the outlined acceleration, one can find ϵ-solution to the problem in*

$$M(\epsilon) = O(1) R \|P\|_{1,2} \sqrt{\ln(n)/\epsilon},$$
$$\|P\|_{1,r} = \max_j \|\mathrm{Column}_j(P)\|_r$$

steps, with effort per step dominated by two matrix-vector multiplications involving P and P^T.

♣ **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

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

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

♣ 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:**

- ζ 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* ζ *is simple:* given u, it takes $O(1)(p+q)$ 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 Π_y on E_y such that $\mathbf{E}_{\eta\sim\Pi_y}\{\eta\}\equiv y$.

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

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

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

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

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

$\Rightarrow$ ....

$\Rightarrow$ $\{\xi_x,\, x\in X\};\ \{\eta_y, y\in Y\};\ \omega(x,y):Z:=X\times Y\to\mathbf{R};\ \Omega_x,\Omega_y,\sigma^2$

Randomized MP Algorithm

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

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

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

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

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

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

$\Rightarrow$ .... $\Rightarrow \Omega_x, \Omega_y, \sigma$

Theorem [Ioud.&Nem.'11]

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

$$\mathbf{E}\left\{f(x^N) - \mathrm{Opt}\right\} \le 7\max\left[\frac{\sigma}{\sqrt{N}}, \frac{\|A\|_{x,y}\sqrt{\Omega_x\Omega_y}}{N}\right].$$

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

$$\mathbf{E}\left\{\varepsilon_{\mathrm{sad}}(z^N)\right\} \le 7\max\left[\frac{\sigma}{\sqrt{N}}, \frac{\|A\|_{x,y}\sqrt{\Omega_x\Omega_y}}{N}\right].$$

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., ℓ_1 minimization problems

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

♣ ℓ_1 minimization with uniform fit

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

reduces to a small series of problems

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

Corollary of Theorem:

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

$$\mathrm{Prob}\left\{ \varepsilon_{\mathrm{sad}}(x^N, y^N) \le O(1)\frac{\ln(2mn)\|A\|_{1,\infty}}{\sqrt{N}} \right\} > \frac{1}{2}$$

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 $\epsilon < \|A\|_{1,\infty} = \max_{i,j} |A_{ij}|$ be given. Applying Randomized MP, we with confidence $\geq 1-\beta$ find a feasible solution $(\bar{x}, \bar{y})$ satisfying $\varepsilon_{\rm sad}(\bar{x}, \bar{y}) \leq \epsilon$ in

$$O(1) \ln^2(2mn) \ln(1/\beta)(m+n) \left[\frac{\|A\|_{1,\infty}}{\epsilon}\right]^2$$

arithmetic operations.

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

$$O(1)\sqrt{\ln(2m)\ln(2n)}mn \left[\frac{\|A\|_{1,\infty}}{\epsilon}\right]$$

arithmetic operations.

⇒*When the relative accuracy $\epsilon/\|A\|_{1,\infty}$ 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.*

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

♣ The efficiency estimate

$$O(1)\ln^2(2mn)\ln(1/\beta)(m+n)\left[\frac{\|A\|_{1,\infty}}{\epsilon}\right]^2 \text{ a.o.}$$

says that *with ϵ, β fixed and m, n large, the Randomized MP exhibits sublinear time behavior: ϵ-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 (!).

♣ **Note:** *Similar results hold true for ℓ_1 minimization with 2-fit:*

$$\min_\xi \{\|\xi\|_1 : \|A\xi - b\|_2 \leq \delta\}$$

Numerical Illustration: Policeman vs. Burglar

♣ **Problem:** There are n houses in a city, i-th with wealth w_i. Every evening, Burglar chooses a house i to be attacked, and Policeman chooses his post near a house j. The probability for Policeman to catch Burglar is

$\exp\{-\theta \mathrm{dist}(i,j)\}$, $\mathrm{dist}(i,j)$: distance between houses i and j.

Burglar wants to maximize his expected profit

$$w_i(1 - \exp\{-\theta \mathrm{dist}(i,j)\}),$$

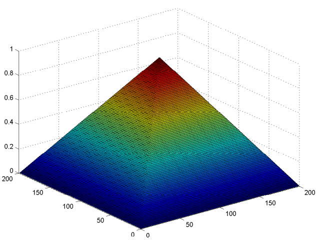
the interest of Policeman is completely opposite.

- *What are the optimal mixed strategies of Burglar and Policeman?*

♠ **Equivalently:** *Solve the matrix game*

$$\max_{\substack{y \geq 0, \\ \sum_{i=1}^n y_i = 1}} \min_{\substack{x \geq 0, \\ \sum_{j=1}^n x_j = 1}} \phi(x,y) := y^T A x$$

$$A_{ij} = w_i(1 - \exp\{-\theta \mathrm{dist}(i,j)\})$$

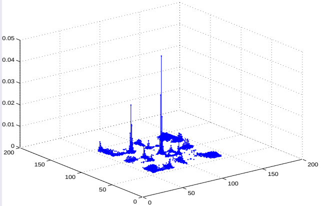


Wealth on 200×200 square grid of houses

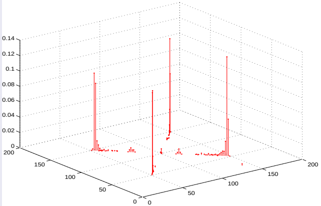
♠ **Deterministic approach:** The 40,000×40,000 fully dense game matrix A is too large for 8 GB RAM of my computer. To compute once $\nabla\phi(x,y) = [A^T y; Ax]$ via on-the-fly generating rows and columns of A takes 97.5 sec (2.67 GHz Intel Core i7 64-bit CPU).

⇒*Running time of Deterministic algorithm is tens of hours...*

♠ **Randomization:** 50,000 iterations of the randomized MP take $1^h 31' 30''$ (like just 28 steps of deterministic algorithm) and result in approximate solution of accuracy 0.0008.

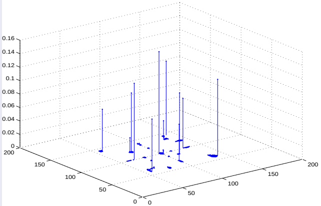


Policeman

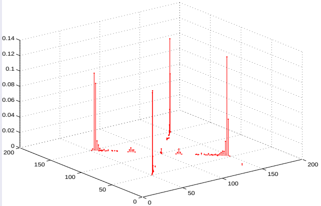


Burglar

♠ The resulting highly sparse near-optimal solution can be refined by further optimizing it on its support by an interior point method. This reduces inaccuracy from 0.0008 to 0.0005 in just 39′.



Policeman, refined



Burglar, refined

- 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, ℓ_1 Minimization... ('10), http://www.optimization-online.org
- A. Juditsky, A. Nemirovski, First Order... I,II: S. Sra, S. Novozin, S.J. Wright, Eds., *Optimization for Machine Learning*, MIT Press, 2011
- 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