Acceleration by Randomization: Randomized First Order Algorithms for Large-Scale Convex Optimization

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

Joint research with Anatoli Juditsky[†], Guanghui Lan[‡], and Alexander Shapiro[§]
[†]: Joseph Fourier University, Grenoble, France; [‡]: ISyE, University of Florida; [§]: ISyE, Georgia Tech

ISMP 2009
Chicago August 23-28, 2009

Claims

Claim I

♣ *When solving large-scale well-structured convex optimization problems*, e.g., the ℓ_1 minimization problem

$$\min_x \{\|x\|_1 : \|Ax - b\|_\infty \le \delta\}$$

with large-scale and possibly dense matrix A,

• *polynomial time methods, and thus generating high accuracy solutions, become prohibitively time consuming;*

• *when medium accuracy solutions are sought, our best choice is computationally cheap first order methods.*

Claim II

♣ *There are interesting and important cases when first order methods can be accelerated significantly by randomization – passing from computationally demanding deterministic first order oracles to their computationally cheap stochastic counterparts.*

Unifying framework: Stochastic Monotone Variational Inequalities

Stochastic Monotone Variational Inequality

$$\text{Find } z_* \in Z : \langle F(z), z - z_* \rangle \geq 0 \, \forall z \in Z$$

$$\left[\begin{array}{l} \bullet\ Z\text{: convex compact in Euclidean space } E \\ \bullet\ F : Z \to E\text{: monotone:} \\ \qquad \langle F(z) - F(z'), z - z' \rangle \geq 0 \ \forall z, z' \in Z \end{array}\right]$$

♠ F is given by *Stochastic Oracle*: at i-th call to SO, $z_i \in Z$ being the input, SO returns $G(z_i, \xi_i) \in E$, where $\xi_1, \xi_2, \ldots$ is a sequence of iid "oracle noises" and

$$\forall z \in Z : \mathbf{E}\{G(z,\xi)\} = F(z) \ \& \ \mathbf{E}\{\|G(z,\xi)\|_*^2\} \leq M^2 < \infty$$

$[\|\cdot\|, \|\cdot\|_*$ – conjugate pair of norms on $E]$

Unifying framework (continued)

$$\text{Find } z_* \in Z : \langle F(z), z - z_* \rangle \geq 0 \,\forall z \in Z \qquad \text{(VI)}$$

♣ Example: Saddle Point case

Let $\phi(x, y) : X \times Y \to \mathbb{R}$ be a convex-concave Lipschitz continuous function, X, Y be convex compact sets.
The saddle points of ϕ on $X \times Y$ are exactly the solutions to (VI) where $Z = X \times Y$ and

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

♠ Accuracy measure for (VI):

- **General case**: $\mathrm{ErrVI}(z) := \sup\limits_{w \in Z} \langle F(w), z - w \rangle$.
- **Saddle Point case:** $\mathrm{ErrSad}(x, y) = \max\limits_{v \in Y} \phi(x, v) - \min\limits_{u \in X} \phi(u, y)$.

Robust Mirror Descent Stochastic Approximation [Nem.&Yudin '79;Jud.&Lan&Nem.&Shap. '08]

Find $z_* \in Z \subset E : \langle F(z), z - z_* \rangle \geq 0 \, \forall z \in Z$ (VI)

♣ Setup:

A norm $\|\cdot\|$ on E and a convex continuous function $\omega : Z \to \mathbb{R}$ such that the set $Z^o = \{z \in Z : \partial\omega(z) \neq \emptyset\}$ is convex, and ω is C^1 and strongly continuous on Z^o:

$$\forall z, z' \in Z^o : \langle \omega'(z) - \omega'(z'), z - z' \rangle \geq \alpha \|z - z'\|^2 \quad [\alpha > 0]$$

♣ The algorithm:

$$\begin{array}{rcl} z_1 & = & z_\omega := \mathrm{argmin}_Z\, \omega(\cdot) \\ z_{t+1} & = & \mathrm{argmin}_Z \left[\langle \gamma G(z_t, \xi_t) - \omega'(z_t), z \rangle + \omega(z)\right] \\ w^N & = & \frac{1}{N} \sum_{t=1}^N z_t \end{array}$$

[• $\gamma > 0$: stepsize • N: # of steps]

Main results on MDSA

$$\text{Find } z_* \in Z \subset E : \langle F(z), z - z_* \rangle \geq 0 \,\forall z \in Z$$
$$G(z,\xi) : \mathbf{E}\{G(z,\xi)\} \equiv F(z) \;\&\; \mathbf{E}\{\|G(z,\xi)\|_*^2\} \leq M^2 \qquad \text{(VI)}$$

Theorem: Let $N \geq 1$. As applied to a monotone SVI, N-step MDSA with appropriately chosen stepsize γ ensures that

$$\mathbf{E}\left\{\mathrm{ErrVI}(w^N)\right\} \leq 2\sqrt{5}\Theta M N^{-\frac{1}{2}}$$
$$\left[\Theta = \sqrt{2\max_{z\in Z}[\omega(z) - \omega(z_\omega) - \langle \omega'(z_\omega), z - z_\omega\rangle]/\alpha}\right]$$

• $\mathbf{E}\left\{\exp\{\|G(z,\xi)\|_*^2/M^2\}\right\} \leq \exp\{1\} \;\forall z \in Z \Rightarrow$
$$\forall \Omega > 1 : \mathrm{Prob}\left\{\mathrm{ErrVI}(w^N) > (8+2\Omega)\sqrt{5}\Theta M N^{-\frac{1}{2}}\right\} \leq 2\exp\{-\Omega\}$$

• $\|G(z,\xi)\|_* \leq M \;\forall z \in Z$ and $\|\cdot\|$-diameter of Z is $\leq D \Rightarrow$
$$\forall \Omega > 0 : \mathrm{Prob}\left\{\mathrm{ErrVI}(w^N) > [2\sqrt{5}\Theta + 5\Omega D] M N^{-\frac{1}{2}}\right\} \leq \exp\{-\tfrac{\Omega^2}{2}\}$$

• In the saddle point case, the above bounds hold true for $\mathrm{ErrSad}(w^N)$ as well.

Good Geometry case

♣ Let $\mathrm{rint}\, Z \subset \mathrm{rint}\, Z^+$, where Z^+ is the direct product of K "simple sets":

- K_b unit Euclidean balls $B_i \subset E_i = \mathbb{R}^{n_i}$;
- K_s spectahedrons $S_j \subset F_j$:

F_j: the space of symmetric matrices of a given block-diagonal structure with the Frobenius inner product

S_j: the set of all positive semidefinite matrices from F_j with unit trace

The norm on $E = \prod_i E_i \times \prod_j F_j \supset Z^+ = \prod_i B_i \times \prod_j S_j$:

We equip E_i with the Euclidean norms $\|\cdot\|_{E_i}$, F_j – with the nuclear norms $\|\cdot\|_{F_j}$, and E – with the norm

$$\|\{e^i, f^j\}\| = \sqrt{\sum_i \|e^i\|_{E_i}^2 + \sum_j \|f^j\|_{F_j}^2}.$$

Good Geometry case (continued)

• With appropriately chosen $\omega(\cdot) : Z \to \mathbb{R}$, we have

$$\Theta = O(1)\sqrt{K \ln\left(\sum_j \dim F_j\right)}$$

$\Rightarrow$*the rate of convergence is nearly dimension-independent:*

$$\mathbf{E}\{\mathrm{Err}(w^N)\} \leq O(1)\sqrt{K \ln\left(\sum_j \dim F_j\right)}M/\sqrt{N}.$$

• With this $\omega(\cdot)$, the per step computational effort is dominated by the processing a single query by the SO plus the cost $\mathcal{C}$ of a prox-step $g \mapsto \mathrm{argmin}_{z\in Z}\left[g^T z + \omega(z)\right]$.

Note: $Z = Z^+ \Rightarrow \mathcal{C} = O(1)\left[\sum_i \dim(E_i) + \sum_j \mathcal{C}_j\right]$

$\mathcal{C}_j$: the cost of eigenvalue decomposition of $A \in F_j$.

$\Rightarrow \mathcal{C} = O(1)\dim Z$ when $Z = Z^+$ and all S_j are simplexes.

Linear case

$$\text{Find } z_* \in Z : \langle F(z), z - z_* \rangle \geq 0 \; \forall z \in Z \qquad \text{(VI)}$$

♣ Let F be affine and monotone:

$$F(z) = \mathcal{A}z + b \;\&\; z^T \mathcal{A} z \geq 0 \;\forall z,$$

and let $Z = \Delta^1 \times ... \times \Delta^K$, $\Delta^j = \{x \in \mathbb{R}^{n_j} : x \geq 0, \sum_i x_i = 1\}$.

Stochastic Oracle

We equip (VI) with the stochastic oracle as follows:

In order to compute $\mathcal{A}z$ *for a given* $z = [z_1; ...; z_K] \in Z$, *we*

- *treat every block* $z_j \in \Delta^j$ *as a probability distribution on the corresponding subset* $I_j \subset \{1, ..., n := \sum_j n_j\}$,
- *pick at random an index* $\imath_j$ *from* I_j, *and*
- *return the sum* $b + \sum_{j=1}^K \mathcal{A}_{:,\imath_j}$ *of* b *and the picked columns* $\mathcal{A}_{:,\imath_j}$ *of* $\mathcal{A}$.

Linear case (continued)

$$\text{Find } z_* \in Z : \langle \mathcal{A}z + b, z - z_* \rangle \geq 0 \, \forall z \in Z \qquad \text{(VI)}$$
$$Z = \Delta^1 \times \ldots \times \Delta^K$$

Note: The output of the above SO can be thought of as a deterministic function $G(z, \xi)$ of z and a random number $\xi \sim \text{Uniform}[0, 1]$. We clearly have

$$\mathbf{E}\{G(z, \xi)\} = F(z) \;\&\; \|G(z, \xi)\|_* \leq M = K^{\frac{1}{2}} \|[\mathcal{A}, b]\|_\infty,$$
$$\|A\|_\infty = \max_{i,j} |A_{ij}|$$

• By the above, for every $\epsilon > 0$, $\beta \in (0, 1)$, MDSA in $N = O(1)K^2 \ln(n/\beta)\|[\mathcal{A}, b]\|_\infty^2 \epsilon^{-2}$ steps generates, with confidence $\geq 1 - \beta$, an ϵ-solution w^N to (VI): $\mathrm{ErrVI}(w^N) \leq \epsilon$.
• A step reduces to extracting from $\mathcal{A}$ K randomly chosen columns plus $O(1)Kn$-a.o. overhead.

• In the saddle point case, ErrVI can be replaced with ErrSad.

Example I: Matrix Game

♣ Matrix game $\min_{x\in\Delta^1}\max_{y\in\Delta^2} y^TAx$ reduces to Linear Monotone VI with the domain $Z=\Delta^1\times\Delta^2$ and the operator

$$F(x,y)=\left[\begin{array}{cc} & A^T\\ -A & \end{array}\right][x;y].$$

• By above, *we can find, with confidence* $\geq 1-\beta$, *an* ϵ*-solution* w^N *to the game:* $\mathrm{ErrSad}(w^N)\leq\epsilon$ *in* $O(1)\ln(n/\beta)(\|A\|_\infty/\epsilon)^2$ *steps,* $n=\dim\Delta^1+\dim\Delta^2$, *a step reducing to extracting from* A *randomly chosen row and column plus an* $O(n)$*-a.o. overhead.*
Note: To build an ϵ-solution, $N=O(1)n\ln(n/\beta)(\|A\|_\infty/\epsilon)^2$ randomly chosen entries in A are inspected. *With* ϵ, β *fixed,* $\dim x=O(\dim y)$ *and* n *large,* N *is incomparably less than the total number* $O(n^2)$ *of data entries.*
• The algorithm has the same performance as the sublinear time randomized game algorithm of Grigoriadis and Khachiyan (1995), and with appropriate choice of $\omega(\cdot)$ becomes nearly identical to the latter.

Application: ℓ_1-minimization

♣ Numerous sparsity-oriented Signal Processing problems reduce to (small series of) problems

$$\min_u \{\|Au - b\|_p : \|u\|_1 \leq 1\} \quad [A : m \times n] \qquad (\ell_1)$$

♠ When $p = \infty$, (ℓ_1) reduces to Matrix Game

$$\min_{x=[u;v]\in\Delta_{2n}} \max_{y=[p;q]\in\Delta_{2m}} [p-q]^T \left[A - b\mathbf{1}^T, -A - b\mathbf{1}^T\right][u;v]$$

$\Rightarrow$MDSA (same as Gr.-Kh. algorithm), can be used to solve ℓ_1-minimization problems.

• When m, n are really large (like 10^4 and more) and A is a general-type dense analytically given matrix, (ℓ_1) becomes prohibitively difficult for all known deterministic algorithms (Interior Point methods, advanced first order algorithms like Smoothing or Mirror Prox,...), and randomized algorithms become a kind of "last resort".

How it works

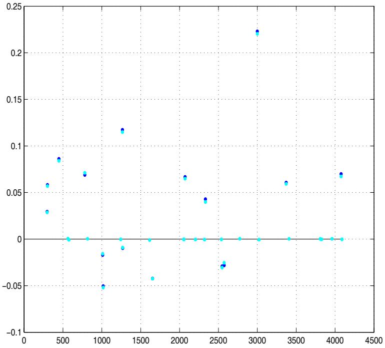
$$\text{Opt} = \min_u \{\|Au - b\|_\infty : \|u\|_1 \leq 1\}$$

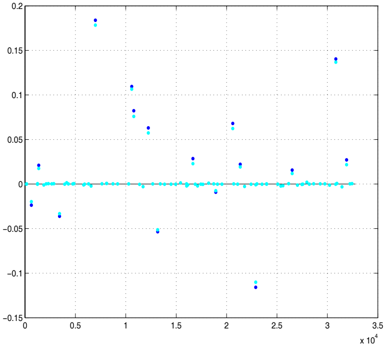
$$\left[\begin{array}{l} A\text{: dense analytically given } m \times n \text{ matrix} \\ \text{b: } \|Au_* - b\|_\infty \leq \delta = 1.\text{e-3 with 16-sparse } u_*,\ \|u_*\|_1 = 1 \end{array}\right]$$

		Errors			CPU	
$m \times n$	Method	$\|u_* - \tilde{u}\|_1$	$\|u_* - \tilde{u}\|_2$	$\|u_* - \tilde{u}\|_\infty$	sec	Mult
2048×4096	MP	0.0014	0.00052	0.00036	122.8	1770
	MDSA	0.039	0.0079	0.0030	325.4	29.3
8192×32768	MP	1.006	0.319	0.184	3141.9	5
	MDSA	0.120	0.0196	0.00634	3000.5	4.7

- MP: Mirror Prox
- $\tilde{u}$: resulting solution
- Last column: (equivalent) # of matrix-vector multiplications.

How it works (continued)





2048×4096 8192×32768

ℓ_1-recovery: Blue: true signal, Cyan: MDSA

Extension: Bilinear Nash Equilibrium

♣ Consider K players, i-th choosing a mixed strategy z_i in the standard simplex $\Delta^i \subset \mathbb{R}^{n_i}$. The loss of i-th player is

$$\phi_i(z_1, ..., z_K) = b_i^T z_i + \sum_j z_i^T A^{ij} z_j$$
$$A^{ij} = -[A^{ji}]^T \forall i \neq j \;\&\; A^{ii} = [A^{ii}]^T \succeq 0 \;\forall i$$

Each player wants to minimize his loss, and we want to find a Nash equilibrium.

♣ This is a particular case of convex Nash Equilibrium problem and thus it reduces to a monotone VI. In our case the domain Z of the VI is $\Delta^1 \times ... \times \Delta^K$, and the operator is linear:

$$F(z_1, ..., z_K) = \mathcal{A}z + [b_1; ...; b_K],\ \mathcal{A} = \begin{bmatrix} 2A^{11} & A^{12} & ... & A^{1K} \\ A^{21} & 2A^{22} & ... & A^{2K} \\ \vdots & \vdots & \ddots & \vdots \\ A^{K1} & A^{K2} & ... & 2A^{KK} \end{bmatrix}$$

Bilinear Nash Equilibrium (continued)

• Here the above reads as follows: *For every* $\epsilon > 0$ *and* $\beta \in (0,1)$*, MDSA, with confidence* $\geq 1-\beta$*, finds* ϵ*-Nash equilibrium* w^N*:*

$$\sum_{i=1}^{K}\left[\phi_i(w^N) - \min_{\zeta_i \in \Delta^i} \phi_i(w_1^N, ..., w_{i-1}^N, \zeta_i, w_{i+1}^N, ..., w_K^N)\right] \leq \epsilon$$

in $N = O(1)K^2 \ln(n/\beta)\|[\mathcal{A}, b]\|_\infty^2 \epsilon^{-2}$ *steps,* $n = \sum_i \dim \Delta_i$*.*
A step reduces to extracting from $\mathcal{A}$ K *randomly chosen columns plus* $O(1)Kn$*-a.o. overhead.*

• For ϵ, β, K and (an upper bound on) $\|[\mathcal{A}, b]\|_\infty$ fixed and $n \to \infty$ this is a sublinear time algorithm.

• When $K = 1$, our problem becomes to minimize over Δ_n a convex quadratic form $z^T Az + b^T z$. Thus, *under normalization* $\|[A, b]\|_\infty \leq 1$*, convex quadratic minimization over the standard simplex (as well as over the unit* $\|\cdot\|_1$*-ball) admits a sublinear time randomized approximation algorithm.*

Example II: Minimizing the maximum of convex polynomials over a simplex/ℓ_1-ball

♣ Consider the optimization problem

$$\min_{x\in\Delta_n}\max_{1\leq j\leq m} p_j(x),\quad p_j(x)=\sum_{\ell=0}^d A_{j\ell}[x,...,x] \qquad (*)$$

• $A_{j\ell}[x_1,...,x_\ell]$: symmetric ℓ-linear forms; • $p_j(\cdot)$: convex.

♠ (*) reduces to the convex-concave saddle point problem

$$\min_{x\in\Delta_n}\max_{y\in\Delta_m}\sum_j y_j p_j(x)$$

and thus – to a monotone VI with $Z=\Delta_n\times\Delta_m$ and

$$F(x,y)=\left[\begin{array}{l} F_x=\sum_{j=1}^m y_j\sum_{\ell=1}^d \ell\{A_{j\ell}[x,...,x,e^\nu]\}_{\nu=1}^n \\ F_y=\{-\sum_{\ell=0}^d A_{\mu\ell}[x,...,x]\}_{\mu=1}^m \end{array}\right]$$

• $e^1,...,e^n$: basic orths.

• Cheap SO for F: $\left[\begin{array}{l} G_x=\sum_{\ell=1}^d \ell\left\{A_{j\ell}[e^{i_1},...,e^{i_{\ell-1}},e^\nu]\right\}_{\nu=1}^n \\ G_y=\left\{-\sum_{\ell=0}^d A_{\mu\ell}[e^{i_1},...,e^{i_\ell}]\right\}_{\mu=1}^m \end{array}\right]$

• $\{1,...,m\}\ni j\sim y$; • $\{1,...,n\}\ni i_1\sim x,...,i_d\sim x$

Minimizing maximum of convex polynomials over a simplex/ℓ_1-ball (continued)

$$\min_{x\in\Delta_n}\max_{1\le j\le m} p_j(x),\quad p_j(x)=\sum_{\ell=0}^d A_{j\ell}[x,...,x] \qquad (*)$$

$$\Rightarrow \qquad G=\left[\begin{array}{l} G_x=\sum_{\ell=1}^d \ell\left\{A_{j\ell}[e^{i_1},...,e^{i_{\ell-1}},e^{\nu}]\right\}_{\nu=1}^n \\ G_y=\left\{-\sum_{\ell=0}^d A_{\mu\ell}[e^{i_1},...,e^{i_\ell}]\right\}_{\mu=1}^m \end{array}\right]$$

$\mathcal{A}$:=maximal magnitude of coefficients in $A_{j\ell}[\cdot,...,\cdot]$

• With this Stochastic Oracle, MDSA with confidence $\geq 1-\beta$ solves $(*)$ within accuracy ϵ in

$$N=O(1)\ln((m+n)/\beta)d^2(\mathcal{A}/\epsilon)^2$$

calls to the oracle, with $O(m+dn)$-a.o. overhead per step.

• A call to the oracle reduces to extracting $O(1)d(m+n)$ coefficients of the forms $A_{j\ell}[\cdot,\cdot,...,\cdot]$, given the "addresses" $j,\ell,i_1,...,i_\ell$ of the coefficients.

Example III: Least Squares ℓ_1-minimization

$$\min_u \{\|Au-b\|_p : \|u\|_1 \le 1\} \quad [A: m\times n] \qquad (\ell_1)$$

♣ When $p=\infty$, (ℓ_1) reduces to Matrix Game. What happens in the Least Squares case $p=2$?

♠ (ℓ_1) reduces to the saddle point problem

$$\min_{x\in\Delta_{2n}} \max_{y\in B_m} y^TCx, \quad C=[A-b\mathbf{1}^T, -A-b\mathbf{1}^T]$$

and thus – to a monotone VI with $Z=\Delta_{2n}\times B_m$ and

$$F(x,y)=[F_x(y)=C^Ty; F_y(x)=-Cx].$$

♠ F can be represented by a Stochastic Oracle:

$$G_x^T=\|y\|_1\mathrm{sign}(y_\imath)C_{\imath,:}, \; G_y=C_{:,\jmath}$$
$$\mathrm{Prob}\{\imath=i\}=|y_i|/\|y\|_1, \; \mathrm{Prob}\{\jmath=j\}=x_j$$

• The associated norm is $\|[x;y]\|=\sqrt{\|x\|_1^2+\|y\|_2^2}$

$$\Rightarrow \quad M=\sqrt{2}\max\left[\|C\|_{1,2}, \sqrt{m}\|C\|_\infty\right]$$

• $\|C\|_{1,2}=\max_j\|C_{:,j}\|_2$ • $\|C\|_\infty=\max_{i,j}|C_{ij}|$

Least Squares ℓ_1-minimization (continued)

$$\min_u \{\|Au-b\|_2 : \|u\|_1 \le 1\} \quad [A : m\times n]$$
$$\Rightarrow C = [A - b\mathbf{1}^T, -A - b\mathbf{1}^T] \qquad (\ell_1)$$
$$\Rightarrow M = \sqrt{2}\max\left[\|C\|_{1,2}, \sqrt{m}\|C\|_\infty\right]$$

• MDSA with confidence $\ge 1-\beta$ finds ϵ-solution to (ℓ_1) in $O(1)\ln(mn/\beta)(M/\epsilon)^2$ steps reducing to extracting from $[A,b]$ randomly chosen row and column plus $O(m+n)$-a.o. overhead.

Difficulty: The "true" scale parameter in our problem is $\|C\|_{1,2}$; and M can be greater than $\|C\|_{1,2}$ by "large" factor $\sqrt{m}$

Remedy: Randomized preprocessing $[A,b]\mapsto UD[A,b]$ with orthogonal U, $|U_{ij}|\le O(m^{-1/2})$ and random diagonal D with iid $D_{ii}\sim \text{Uniform}\{-1;1\}$ results in an equivalent problem, preserves $\|C\|_{1,2}$ and with confidence $\ge 1-\beta$ makes M "small": $M \le O(1)\sqrt{\ln(mn/\beta)}\|C\|_{1,2}$.
• With properly chosen U (e.g., Hadamard or DFT matrix), the preprocessing costs just $O(1)mn\ln(m)$ a.o.

Stochastic Mirror Prox algorithm

$$\text{Find } z_* \in Z : \langle F(z), z - z_* \rangle \geq 0 \, \forall z \in Z$$
$$G(z,\xi) : \mathbf{E}\{G(z,\xi)\} = F(z), \ \mathbf{E}\{\|G(z,\xi)\|_*^2\} \leq M^2$$
$$\overset{\text{MDSA}}{\Rightarrow} \mathbf{E}\{\mathrm{Err}(w^N)\} \leq O(1)\Theta M/\sqrt{N}$$

Bad news: It does not matter whether M comes from noise or from F: no acceleration when passing from *noisy* observations of *nonsmooth* F to *precise* observations of *smooth* F.

Assumption:

$$\|F(z) - F(z')\|_* \leq L\|z - z'\| + M$$
$$\mathbf{E}\{G(z,\xi)\} \equiv F(z), \ \mathbf{E}\{\|G(z,\xi) - F(z)\|_*^2\} \leq M^2$$

SMP algorithm:

$$\begin{aligned}
w_t &= \mathrm{argmin}_Z \{\langle \gamma G(z_t, \xi_{2t-1}) - \omega'(z_{t-1}), z\rangle + \omega(z)\} \\
z_{t+1} &= \mathrm{argmin}_Z \{\langle \gamma G(w_t, \xi_{2t}) - \omega'(z_{t-1}), z\rangle + \omega(z)\} \\
w^N &= \frac{1}{N}\sum_{t=1}^N w_t
\end{aligned}$$

Main results on SMP

$$\begin{array}{c}\text{Find } z_* \in Z \subset E : \langle F(z), z - z_*\rangle \geq 0\, \forall z \in Z \\ F: \text{ monotone}, \|F(z) - F(z')\|_* \leq L\|z - z'\| + M \\ G(z,\xi): \mathbf{E}\{G(z,\xi)\} \equiv F(z) \;\&\; \mathbf{E}\{\|G(z,\xi) - F(z)\|_*^2\} \leq M^2\end{array} \qquad \text{(VI)}$$

Theorem: Let $N \geq 1$. As applied to a monotone SVI, N-step SMP with appropriately chosen stepsize γ ensures that

$$\mathbf{E}\left\{\mathrm{ErrVI}(w^N)\right\} \leq 2\Theta\left[\Theta L N^{-1} + 4MN^{-\frac{1}{2}}\right]$$

$\left[\Theta\text{: given by setup, } \leq O(1)\sqrt{K\ln(\dim Z)} \text{ in the Good Geometry case}\right]$

- When $\mathbf{E}\{\exp\{\|G(z,\xi) - F(z)\|_*^2/M^2\}\} \leq \exp\{1\}$, one has $\forall \Omega > 0$:

$$\begin{array}{r}\mathrm{Prob}\left\{\mathrm{ErrVI}(w^N) > 2\Theta\left[\Theta L N^{-1} + 4MN^{-\frac{1}{2}}\right] + 4\Omega\Theta M N^{-\frac{1}{2}}\right\} \\ \leq \exp\{-\Omega^2/3\} + \exp\{-\Omega N\}\end{array}$$

- In the Saddle Point case, the bounds are valid for $\mathrm{ErrSad}(w^N)$ as well.

$$\boxed{\begin{array}{c}\text{Find } z_* \in Z \subset E : \langle F(z), z - z_*\rangle \geq 0\, \forall z \in Z\\ F:\ \text{monotone}, \|F(z) - F(z')\|_* \leq L\|z - z'\|\\ G(z,\xi) : \mathbf{E}\{G(z,\xi)\} \equiv F(z)\ \&\ \mathbf{E}\{\|G(z,\xi) - F(z)\|_*^2\} \leq M^2\end{array}}$$

$$\Downarrow$$

$$\mathbf{E}\left\{\mathrm{Err}(w^N)\right\} \leq 2\Theta\left[\Theta L N^{-1} + 4MN^{-\frac{1}{2}}\right]$$

♣ Let a Lipschitz continuous operator F be represented by a cheap Stochastic Oracle.

By calling the oracle several times and averaging the answers, we reduce the "level of noise" M, and thus reduce, to some extent, the number of steps required to reach a desired accuracy.

This added flexibility can be instrumental in the case of expensive prox-steps $g \mapsto \mathrm{argmin}_z\{\langle g, z\rangle + \omega(z)\}$.

Example: Eigenvalue minimization

♣ Consider Eigenvalue Minimization problem

$$\min_{x \in \Delta_n} f(x) := \lambda_{\max}\left(\sum_{i=1}^n x_i A_i\right)$$

- $A_i \in \mathbf{S}^n$: spectral norm ≤ 1, at most S nonzero entries

(P)

♠ We want to solve the problem within fixed accuracy $\epsilon \ll 1$ in the case where
- n, m are large
- A_i are sparse ($S \ll m^2$), while $\sum_i y_i A_i$ can be dense

Note: Below, $\approx | \gtrsim$ means "$= | \geq$ *up to logarithmic in* m, n, ϵ^{-1} *factors*".

Eigenvalue minimization (continued)

$$\min_{x\in\Delta_n} f(x) \equiv \lambda_{\max}\left(\sum_{i=1}^n x_i A_i\right)$$

- $A_i \in \mathbf{S}^n$: spectral norm ≤ 1, at most S nonzero entries
- $\Delta_n = \{x \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$, $\epsilon \ll 1$ fixed, $n, m \to \infty$

(P)

♠ When $n \gg m$, the best complexity of solving (P) by fully deterministic algorithms is $\mathcal{C}_{\rm d} \approx \epsilon^{-1}\left[m^3 + nS\right]$ a.o.

• achievable with Smoothing (Nesterov '03) or Deterministic Mirror Prox (Nem. '04) algorithms as applied to the Saddle Point reformulation of (P):

$$\min_{x\in\Delta_n}\max_{y\in S_m} \mathrm{Tr}\left(y[\sum_{i=1}^n x_i A_i]\right)$$
$$S_m = \left\{y \in \mathbf{S}^m_+ : \mathrm{Tr}(y) = 1\right\} \qquad \text{(SP)}$$

Eigenvalue minimization (continued)

$$\min_{x\in\Delta_n} f(x) \equiv \lambda_{\max}\left(\sum_{i=1}^n x_i A_i\right)$$

- $A_i \in \mathbf{S}^n$: spectral norm ≤ 1, at most S nonzero entries
- $\Delta_n = \{x \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$, $\epsilon \ll 1$ fixed, $n \gg m \to \infty$

(P)

♠ Another available complexity bound is

$$\mathcal{C}_{\mathrm{r}} \approx \epsilon^{-2}\left[nS + \epsilon^{-1} m^2\right] \text{ a.o.}$$

• achievable with a "slightly randomized" algorithm (ℓ_1-Mirror Descent, with values and subgradients of $f(\cdot)$ approximated by the Power method).

Eigenvalue minimization (continued)

$$\min_{x\in\Delta_n}\max_{y\in S_m}\mathrm{Tr}\left(y[\textstyle\sum_{i=1}^n x_iA_i]\right) \qquad \text{(SP)}$$

♠ Equipping (SP) with properly built Stochastic Oracle and applying SMP, the complexity bound becomes

$$\mathcal{C}_* \approx \epsilon^{-2}\left[nS+\epsilon^{-1}\min[\epsilon^{-2}S+m,m^2]\right] \text{ a.o.}$$

♠ *In a meaningful range of the values of* m,n,ϵ *this bound is by far better than the known alternatives.* E.g. when

$$n=O(m^{1+\kappa}),\ S=O(m^{1-\kappa}),\ \epsilon=O(m^{-\gamma}) \qquad [0\le\kappa\le1]$$

one has

$$\mathcal{C}_{\rm d}/\mathcal{C}_* \gtrsim \begin{cases} m^{1-\gamma}. & 0 \le \gamma \le \frac{1+\kappa}{3}\\ m^{2+\kappa-4\gamma}, & \frac{1+\kappa}{3} \le \gamma \le \frac{2+\kappa}{4}\end{cases}$$

$$\mathcal{C}_{\rm s}/\mathcal{C}_* \gtrsim \begin{cases} m^{\gamma}, & 0 \le \gamma \le \frac{1+\kappa}{3}\\ m^{1+\kappa-2\gamma}, & \frac{1+\kappa}{3} \le \gamma \le \frac{1+\kappa}{2}\end{cases}$$

Eigenvalue minimization (continued)

$$\min_{y\in\Delta_n}\max_{z\in S_m} \mathrm{Tr}\left(z[\sum_{i=1}^n y_i A_i]\right)$$
$$\Downarrow$$
Find $(y_*, z_*) \in Z = \Delta_n \times S_m$:
$$\langle F(x,y), (x,y) - (y_*, z_*)\rangle \geq 0 \ \forall (x,y) \in Z \qquad \text{(VI)}$$
$$F(x,y) = (F_x(y) = [\mathrm{Tr}(yA_1);...;\mathrm{Tr}(yA_n)], F_y(x) = -\sum_i x_i A_i)$$

♠ To build an unbiased estimate G_y of $F_y(x)$, we treat $x \in \Delta_n$ as a probability distribution on $\{1,...,n\}$. $\widehat{F}_y$ is the average of p matrices picked at random from $\{-A_1,...,-A_n\}$.

Eigenvalue minimization (continued)

Find $(x_*, y_*) \in Z = \Delta_n \times S_m$:

$$\langle F(x,y), (x,y) - (x_*, y_*)\rangle \geq 0 \;\forall (x,y) \in Z$$

$$F(x,y) = (F_x(y) = [\mathrm{Tr}(yA_1); \ldots; \mathrm{Tr}(yA_n)], F_y(x) = -\sum_i x_i A_i)$$

♠ To build an estimate G_x of $F_x(y)$, we use *exponential representation* $y = [\mathrm{Tr}(\exp\{2w\})]^{-1}\exp\{2w\}$ of $y \in S_m$.
Note: When running N-step SMP, the required matrices w are readily available, have $\leq NpS$ nonzeros, and $\|w\| \lesssim N$.

♡ The estimate is (cf. Arora & Kale, '07)

$$G_x = \left[\sum_{i=1}^p \zeta_i^T \zeta_i\right]^{-1} \sum_{i=1}^p \left[\zeta_i^T A_1 \zeta_i; \ldots; \zeta_i^T A_n \zeta_i\right]$$
$$\zeta_i = \exp\{w\}\xi_i,\ \xi_i \sim \mathcal{N}(0, I_m)$$

Note: $\exp\{w\}\xi$ is computed to high accuracy as $\sum_{\nu=0}^M \frac{1}{\nu!} w^\nu \xi$