

# Tutorial: *Mirror Descent Algorithms for Large-Scale Deterministic and Stochastic Convex Optimization*

Arkadi Nemirovski

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

COLT 2012  
Edinburgh, June 24-27, 2012

- Basic Mirror Descent
- Favorable Geometry Case
- Stochastic Oracle
- $O(1/t)$  Convergence via ExtraGradient Implementation
- Acceleration by Randomization

### ♣ Problem of Primary Interest: *Convex Minimization*

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

- $X$ : convex compact subset of Euclidean space  $E$
- $f : X \rightarrow \mathbf{R}$ : convex Lipschitz continuous

♠  $f$  is represented by a *First Order oracle*:

- given on input  $x \in X$ , FO returns the value  $f(x)$  and a subgradient  $f'(x)$  of  $f$  at  $x$
- the vector field  $x \mapsto f'(x)$  is assumed to be **bounded** on  $X$

### ♣ Mirror Descent for $(P)$ , *milestones*:

- *Subgradient Descent* (“Euclidean prototype”): N. Shor, 1967:

$$X \ni x_\tau \mapsto x_{\tau+1} = \text{Proj}_X(x_\tau - \gamma_\tau f'(x_\tau))$$

- $\gamma_\tau > 0$ : stepsizes •  $\text{Proj}_X(y) = \text{argmin}_{z \in X} \|y - z\|_2$
- *General Mirror Descent scheme*: Nem., 1979
- *Modern Proximal Point form*: A. Beck & M. Teboulle, 2003

# Proximal Setup

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

- $X$ : convex compact subset of Euclidean space  $E$

♣ **Setup for MD** (“proximal setup”) is given by

- a *norm*  $\| \cdot \|$  on  $E$
- a *distance-generating function*  $\omega(x) : X \rightarrow \mathbf{R}$  which should be

- *convex and continuous on  $X$*
- *strongly convex, modulus 1, w.r.t.  $\| \cdot \|$ :*

$$\langle \omega'(x) - \omega'(x'), x - x' \rangle \geq \|x - x'\|^2$$

for all  $x, x' \in X^\circ = \{x \in X : \partial\omega(x) \neq \emptyset\}$

- admitting a continuous on  $X^\circ$  selection  $\omega'(x)$  of subgradients

♠ **Example:** *Euclidean setup:*

$$E = \mathbf{R}^n, \|x\| = \|x\|_2, \omega(x) = \frac{1}{2}x^T x$$

# Proximal Setup (continued)

♣ **Proximal setup**  $\|\cdot\|, \omega(\cdot)$  for  $X \subset E$  induces:

- $\omega$ -center of  $X$   $x_\omega = \operatorname{argmin}_{x \in X} \omega(x)$
- Bregman distance  $V_x(y) = \omega(y) - \omega(x) - \langle \omega'(x), y - x \rangle$ ,  
 $x \in X^\circ, y \in X$ . By strong convexity of  $\omega(\cdot)$ ,

$$V_x(y) \geq \frac{1}{2} \|y - x\|^2$$

- $\omega$ -radius of  $X$   $\Omega = \Omega[X, \omega(\cdot)] = \sqrt{2[\max_{x \in X} \omega(x) - \min_{x \in X} \omega(x)]}$

For  $x \in X$  one has

$$\frac{1}{2} \|x - x_\omega\|^2 \leq V_{x_\omega}(x) \leq \omega(x) - \omega(x_\omega) \leq \frac{1}{2} \Omega^2$$

$$\Rightarrow \|x - x_\omega\| \leq \Omega \quad \forall x \in X$$

- prox-mapping

$$[x \in X^\circ, \xi \in E] \mapsto \operatorname{Prox}_x(\xi) := \operatorname{argmin}_{z \in X} [\langle \xi, z \rangle + V_x(z)] \in X^\circ$$

♠ With Euclidean setup,

$$V_x(y) = \frac{1}{2} \|x - y\|_2^2, \quad \operatorname{Prox}_x(\xi) = \operatorname{Proj}_X(x - \xi)$$

$\Rightarrow$  Subgradient Descent is the recurrence

$$x_{\tau+1} = \operatorname{Prox}_{x_\tau}(\gamma_\tau f'(x_\tau))$$

# Basic Mirror Descent

- $X$ : convex compact subset of Euclidean space  $E$
- $\|\cdot\|, \omega(\cdot)$ : proximal setup for  $(E, X)$

♣ **MD** works with a *sequence of vector fields*  $\{g_\tau(\cdot) : X \rightarrow E\}_\tau$  represented by *an oracle*. At call  $\tau = 1, 2, \dots$ , the query point being  $x_\tau$ , the oracle returns the vector  $g_\tau(x_\tau) \in E$ .

- In most of applications, the sequence  $\{g_\tau(\cdot)\}_\tau$  is just stationary:  $g_\tau(\cdot) \equiv g(\cdot)$ .

♠ MD is the recurrence

$$x_1 = x_\omega := \operatorname{argmin}_X \omega(\cdot); \quad x_{\tau+1} = \operatorname{Prox}_{x_\tau}(\gamma_\tau g_\tau(x_\tau))$$

- $x_\tau \in X^o$ : search points
- $\gamma_\tau > 0$ : stepsizes

# Basic Mirror Descent (continued)

$$x_1 = x_\omega := \operatorname{argmin}_X \omega; x_{\tau+1} = \operatorname{Prox}_{x_\tau}(\gamma_\tau g_\tau(x_\tau))$$

♣ **Main Property of MD:** *Under Boundedness Assumption*

$$\sup_{x \in X, \tau} \|g_\tau(x)\|_* \leq L < \infty$$

•  $\|\xi\|_* = \max\{\langle \xi, x \rangle : \|x\| \leq 1\}$  is the conjugate of  $\|\cdot\|$   
*the residual*

$\mathcal{E}_t := \max_{z \in X} \sum_{\tau \leq t} \lambda_\tau^t \langle g_\tau(x_\tau), x_\tau - z \rangle$ ,  $\lambda_\tau^t = \gamma_\tau / \sum_{s \leq t} \gamma_s$   
*obeys the bound*

$$\mathcal{E}_t \leq \frac{\Omega^2 + \sum_{\tau \leq t} \gamma_\tau^2 \|g_\tau(x_\tau)\|_*^2}{2 \sum_{\tau \leq t} \gamma_\tau}, t = 1, 2, \dots$$

• In particular, when  $\frac{\Omega}{L\sqrt{t}} \leq \gamma_\tau \leq \frac{\Omega}{\|g_\tau(x_\tau)\|_*\sqrt{t}}$  for  $1 \leq \tau \leq t$  (e.g.,  $\gamma_\tau = \frac{\Omega}{L\sqrt{t}}$ , or  $\gamma_\tau = \frac{\Omega}{\|g_\tau(x_\tau)\|_*\sqrt{t}}$ ,  $1 \leq \tau \leq t$ ), one has

$$\mathcal{E}_t \leq \Omega L / \sqrt{t}.$$

♠ **Fact:** When  $g_\tau(\cdot)$  come from problem “with convex structure,” the residual  $\mathcal{E}_t$  upper-bounds inaccuracy of the approximate solution  $x^t := \sum_{\tau \leq t} \lambda_\tau^t x_\tau$  to the problem.

# Basic Mirror Descent (continued)

**Example 1: Convex Minimization**  $\text{Opt} = \min_X f$ . Applying MD to  $\{g_\tau(\cdot) \equiv f'(\cdot)\}_\tau$  and assuming w.l.o.g. the Lipschitz constant  $L_{\|\cdot\|}(f)$  of  $f$  taken w.r.t.  $\|\cdot\|$  to upper-bound  $\|f'(\cdot)\|_*$ , one has  $f(x^t) - \text{Opt} \leq \mathcal{E}_t$ :

$$\begin{aligned}\mathcal{E}_t &= \max_{z \in X} \sum_{\tau \leq t} \lambda_\tau^t \langle f'(x_\tau), x_\tau - z \rangle \geq \max_{z \in X} \sum_{\tau \leq t} \lambda_\tau^t [f(x_\tau) - f(z)] \\ &\geq \max_{z \in X} [f(\sum_{\tau \leq t} \lambda_\tau^t x_\tau) - f(z)] = f(x^t) - \text{Opt}\end{aligned}$$

$\Rightarrow$  For every  $t$ ,  $t$ -step MD with appropriate stepsizes ensures  $f(x^t) - \text{Opt} \leq \Omega L_{\|\cdot\|}(f) / \sqrt{t}$

**Example 1.A: Convex Online Minimization.** When  $g_\tau(x) = f'_\tau(x)$ , with convex functions  $f_\tau(\cdot) : X \rightarrow \mathbf{R}$  satisfying  $\|f'_\tau(x)\|_* \leq L < \infty$  for all  $x \in X, \tau$ ,  $t$ -step MD with stepsizes  $\gamma_\tau = \frac{\Omega}{L\sqrt{t}}$ ,  $1 \leq \tau \leq t$ , ensures that

$$\frac{1}{t} \sum_{\tau \leq t} f_\tau(x_\tau) \leq \frac{\Omega L}{\sqrt{t}} + \min_{x \in X} \frac{1}{t} \sum_{\tau \leq t} f_\tau(x)$$



# Basic Mirror Descent (continued)

## Example 2: Convex-Concave Saddle Point problem

$$\text{SadVal} = \min_{u \in U} \max_{v \in V} f(u, v).$$

### ♣ Situation:

- $X = U \times V \subset E_U \times E_V =: E$  with compact convex  $U, V$
- $f(u, v) : X \rightarrow \mathbf{R}$ : convex in  $x \in U$ , concave in  $v \in V$ , Lipschitz continuous

♠  $f, U, V$  give rise to two convex optimization problems:

$$\text{Opt}(P) = \min_{u \in U} [\bar{f}(u) := \max_{v \in V} f(u, v)] \quad (P)$$

$$\text{Opt}(D) = \max_{v \in V} [\underline{f}(v) := \min_{u \in U} f(u, v)] \quad (D)$$

with equal optimal values:  $\text{Opt}(P) = \text{Opt}(D)$ , and to *vector field*

$$g(x = [u; v]) = \begin{bmatrix} g_u(u, v) \in \partial_u f(u, v) \\ g_v(u, v) \in \partial_v (-f(u, v)) \end{bmatrix} : X := U \times V \rightarrow E$$

♠ Optimal solutions  $u_*, v_*$  to  $(P), (D)$  are exactly the *saddle points* of  $f$  on  $U \times V$ :

$$f(u, v_*) \geq f(u_*, v_*) \geq f(u_*, v) \quad \forall (u \in U, v \in V) :$$

# MD for Saddle Point problems

$$\text{Opt}(P) = \min_{u \in U} [\bar{f}(u) := \max_{v \in V} f(u, v)] \quad (P)$$

$$\text{Opt}(D) = \max_{v \in V} [\underline{f}(v) := \min_{u \in U} f(u, v)] \quad (D)$$

$$\Rightarrow g(u; v) = [f'_u(u, v); -f'_v(u, v)] : U \times V \rightarrow E$$

♣ **Fact:** Applying MD to  $g_\tau(\cdot) \equiv g(\cdot)$ , the residual

$\mathcal{E}_t = \max_{z \in X} \sum_{\tau \leq t} \lambda_\tau \langle g(x_\tau), x_\tau - z \rangle$ ,  $\lambda_\tau = \gamma_\tau / \sum_{s \leq t} \gamma_s$   
upper-bounds the saddle point inaccuracy (“duality gap”) of the  
approximate solution  $x^t = [u^t; v^t] := \sum_{\tau \leq t} \lambda_\tau x_\tau$  to  $(P, D)$ :

$$[\bar{f}(u^t) - \text{Opt}(P)] + [\text{Opt}(D) - \underline{f}(v^t)] = \bar{f}(u^t) - \underline{f}(v^t) \leq \mathcal{E}_t$$

$$\begin{aligned} \forall [u; v] \in U \times V : \mathcal{E}_t &\geq \sum_{\tau \leq t} \lambda_\tau \langle g(x_\tau), x_\tau - [u; v] \rangle \\ &= \sum_{\tau \leq t} \lambda_\tau [\langle f'_u(u_\tau, v_\tau), u_\tau - u \rangle + \langle -f'_v(u_\tau, v_\tau), v_\tau - v \rangle] \\ &\geq \sum_{\tau \leq t} \lambda_\tau [f(u_\tau, v_\tau) - f(u, v_\tau) - f(u_\tau, v_\tau) + f(u_\tau, v)] \\ &= \sum_{\tau \leq t} \lambda_\tau [f(u_\tau, v) - f(u, v_\tau)] \geq f(u^t, v) - f(u, v^t) \\ &\Rightarrow \mathcal{E}_t \geq \max_{u \in U, v \in V} [f(u^t, v) - f(u, v^t)] = \bar{f}(u^t) - \underline{f}(v^t). \end{aligned}$$

# MD for Saddle Point problems (continued)

$$\text{Opt}(P) = \min_{u \in U} [\bar{f}(u) := \max_{v \in V} f(u, v)] \quad (P)$$

$$\text{Opt}(D) = \max_{v \in V} [\underline{f}(v) := \min_{u \in U} f(u, v)] \quad (D)$$

$$\Rightarrow g(u, v) = [f'_u(u, v); -f'_v(u, v)] : U \times V \rightarrow E$$

♠ Assuming that  $\|\cdot\|$  respects representation  $E = E_U \times E_V$ :  
 $\|[u; v]\| \equiv \|[u; -v]\|$ , we can ensure that  $\|g(\cdot)\|_* \leq L_{\|\cdot\|}(f)$ .  
 $\Rightarrow$  *t-step MD with properly chosen stepsizes ensures that*  
 $[\bar{f}(u^t) - \text{Opt}(P)] + [\text{Opt}(D) - \underline{f}(v^t)] \leq \Omega L_{\|\cdot\|}(f) / \sqrt{t}$ .

♠ Similar results for other “problems with convex structure:”

- variational inequalities with monotone operators
- convex Nash equilibrium problems

# Role of Symmetry

$$\mathcal{E}_t \leq \Omega[\sup_{x \in X, \tau} \|g_\tau(x)\|_*] / \sqrt{t} \quad (*)$$

♣ When  $X$  is “nearly symmetric,” the MD efficiency estimate can be improved. Assume that

- $X$  contains  $\|\cdot\|$ -ball of radius  $\theta\Omega$
- The vector fields  $\{g_\tau(\cdot)\}_\tau$  are uniformly semi-bounded:

$$M := \sup_{x, x' \in X, \tau} \langle g_\tau(x), x' - x \rangle < \infty$$

Then for every  $t \geq 4/\theta^2$ , the  $t$ -step MD with the stepsizes

$$\gamma_\tau = \frac{\Omega}{\|g_\tau(x_\tau)\|_* \sqrt{t}} \quad 1 \leq \tau \leq t$$

ensures that

$$\mathcal{E}_t \leq 2\theta^{-1} M / \sqrt{t} \quad (!)$$

♠ **Note:** When  $\theta = O(1)$ ,

- (!) can only be better than (\*)
- When  $g_\tau(\cdot) \equiv g(\cdot)$  comes from  $\min_{u \in U} \max_{v \in V} f(u, v)$ , we have  $M \leq \max_{U \times V} f - \min_{U \times V} f \Rightarrow$  (!) becomes

$$\mathcal{E}_t \leq O(1) [\max_{U \times V} f - \min_{U \times V} f] / \sqrt{t}$$

# $O(1/\sqrt{t})$ – good or bad?

♣ The MD convergence rate  $O(1/\sqrt{t})$  is slow. However, *this is the best possible rate one can expect when solving nonsmooth large-scale convex problems represented by FO oracles, or any other oracles providing local information.*

♠ **Bad news:** Consider Convex Minimization problem

$$\text{Opt}(f) = \min_x \{f(x) : \|x\| \leq R\} \quad (P_f)$$

where  $\|\cdot\|$  is either the norm  $\|\cdot\|_p$  on  $E = \mathbf{R}^n$  ( $p = 1, 2$ ), or the nuclear norm on  $\mathbf{R}^{n \times n}$ . Let

$$\mathcal{F}_{\|\cdot\|}(L) = \{f : E \rightarrow \mathbf{R} : f \text{ is convex, } L_{\|\cdot\|}(f) \leq L\},$$

and assume that when solving  $(P_f)$ ,  $f \in \mathcal{F}_{\|\cdot\|}(L)$  is learned via calls, one per step, to a FO (or any local) oracle.

Then for every  $t \leq n$  and any  $t$ -step algorithm  $\mathcal{B}$  one has

$$\sup_{f \in \mathcal{F}_{\|\cdot\|}(L)} [f(x_{\mathcal{B}}(f)) - \text{Opt}(f)] \geq 0.01 LR/\sqrt{t}$$

- $x_{\mathcal{B}}(f)$ : solution generated in  $t$  steps by  $\mathcal{B}$  as applied to  $(P_f)$

# $O(1/\sqrt{t})$ – good or bad? (continued)

$\text{Opt}(f) = \min_{x \in X} f(x)$ ,  $X \subset X_R := \{x \in E : \|x\| \leq R\}$  ( $P_f$ )  
 $\|\cdot\|$ :  $\|\cdot\|_p$  norm on  $E = \mathbf{R}^n$  ( $p = 1, 2$ ), or nuclear norm on  $\mathbf{R}^{n \times n}$ .

♠ **Relatively good news:** With appropriate proximal setup,  $t$ -step MD as applied to  $(P_f)$  ensures

$$f(x^t) - \text{Opt}(f) \leq O\left(L_{\|\cdot\|}(f)R/\sqrt{t}\right)$$

- hidden factor:  $O(1)$  when  $\|\cdot\| = \|\cdot\|_2$ , otherwise  $O(1)\sqrt{\ln(n+1)}$

**Note:**

- Rate of convergence is (nearly) *dimension-independent*
- When  $X$  is simple, computational effort per MD step in the large scale case is *by order of magnitudes* smaller than in all known polynomial time Convex Optimization techniques, like Interior Point methods

⇒ When solving problems with convex structure to *low* or *medium* accuracy, MD could be the method of choice...

# Favorable Geometry case

$$\mathcal{E}_t \leq \Omega[X, \omega] \sup_{x \in X, \tau} \|g_\tau(x)\|_* / \sqrt{t}$$

♣ **Question:** *How to choose a good proximal setup?*

- In general, the answer depends on the geometry of  $X$  and on a priori information on  $\{g_\tau(\cdot)\}_\tau$
- There is, however, a *favorable geometry* case when the answer is clear:
  - Assuming w.l.o.g. that  $X^+ = \frac{1}{2}[X - X]$  linearly spans  $E$ ,  $X^+$  is the unit ball of norm  $\|\cdot\|_X$  given solely by  $X$ .
  - A *Favorable Geometry case* is the one where  $X$  admits a d.-g.f.  $\omega_X(\cdot)$  such that  $\|\cdot\|_X, \omega_X(\cdot)$  is a valid proximal setup *with "moderate"  $\Omega_X := \Omega[X, \omega_X]$  ( $O(1)$ , or  $O(1) \ln^{O(1)}(\dim X)$ ).*

# Favorable Geometry case (continued)

$$\mathcal{E}_t \leq \Omega[X, \omega] \sup_{x \in X, \tau} \|g_\tau(x)\|_* / \sqrt{t}$$

♠ **Observation:** Let  $\omega_X(\cdot)$  complement  $\|\cdot\|_X$  to a proximal setup. Then for every proximal setup  $\|\cdot\|, \omega(\cdot)$  for  $X$  and every  $\{g_\tau(\cdot)\}_\tau$  one has

$$\begin{aligned} \sup_{x \in X, \tau} \|g_\tau(x)\|_{X,*} &\leq \Omega[X, \omega] \sup_{x \in X, \tau} \|g_\tau(x)\|_* \quad (!) \\ \Rightarrow \Omega_X \sup_{x \in X, \tau} \|g_\tau(x)\|_{X,*} &\leq \Omega_X \Omega[X, \omega] \sup_{x \in X, \tau} \|g_\tau(x)\|_* \end{aligned}$$

$\Rightarrow$  *Passing from  $\|\cdot\|, \omega(\cdot)$  to  $\|\cdot\|_X, \omega_X(\cdot)$  spoils MD efficiency at worst by factor  $\Omega_X = \Omega[X, \omega_X]$ . Thus, with moderate  $\Omega_X$ , the proximal setup  $\|\cdot\|_X, \omega_X(\cdot)$  is nearly optimal.*



# Favorable Geometry: Examples

## ♠ Examples of Favorable Geometry domains $X$ :

$$X = B^1 \times \dots \times B^K$$

where  $K$  is moderate and  $B^k$  are *favorable geometry atoms*:

- $\ell_1/\ell_2$  balls  $B = \{y = [y^1; \dots; y^n] : \sum_{j=1}^n \|y^j\|_2 \leq 1\}$ :

$$\|y\|_B = \sum_{j=1}^n \|y^j\|_2, \quad \omega_B(y) = O(1) \sqrt{\ln(n+1)} \sum_{j=1}^n \|y^j\|_2^{\vartheta_n}$$
$$\vartheta_n = \min[2, 1 + 1/\ln(n)] \Rightarrow \Omega_B \leq O(1) \sqrt{\ln(n+1)}$$

**Note:**  $n = 1$  gives rise to Euclidean setup for  $\|\cdot\|_2$ -ball.

- Nuclear norm balls  $B = \{y \in \mathbf{R}^{m \times n} : \sum_{j=1}^m \sigma_j(y) \leq 1\}$ ,  $m \leq n$ :

$$\|y\|_B = \sum_{j=1}^m \sigma_j(y), \quad \omega_B(y) = O(1) \sqrt{\ln(m+1)} \sum_{j=1}^m \sigma_j^{\theta_m}(y)$$
$$\theta_m = \min[2, 1 + 1/\ln(2m)] \Rightarrow \Omega_B \leq O(1) \sqrt{\ln(m+1)}$$

- ♠ An induced proximal setup for  $X$  is, e.g.,

$$\|(x_1, \dots, x_K)\| = \max_k \|x_k\|_{B^k}, \quad \omega(x_1, \dots, x_K) = \sum_k \omega_{B^k}(x_k)$$
$$\Rightarrow \Omega_X = \sqrt{\sum_k \Omega_{B^k}^2} \leq O(1) \sqrt{K \ln(\dim X)}$$

- $K = O(1) \Rightarrow$  *Favorable Geometry case*. This remains true if  $X \subset B^1 \times \dots \times B^K$  and  $\|\cdot\|_X$  is within  $O(1)$  factor of  $\|\cdot\|$ .

# Favorable Geometry: Counter-Examples

♠ A domain with *intrinsically bad* geometry is the usual box  $X = \{x \in \mathbf{R}^n : \|x\|_\infty \leq 1\}$ . Here  $\Omega[X, \omega] \geq \sqrt{n}$  for all proximal setups with  $\|\cdot\| = \|\cdot\|_X = \|\cdot\|_\infty$ .

♠ In fact, *large-scale*  $\|\cdot\|_p$ -balls with all  $p > 2$  “are bad:”

Let  $p \geq 2$ . Consider Convex Minimization problem

$$\text{Opt}(f) = \min_x \{f(x) : x \in \mathbf{R}^n, \|x\|_p \leq R\}, \quad (P_f)$$

$$f \in \mathcal{F}_{n,p}(L) = \{f : \mathbf{R}^n \rightarrow \mathbf{R} : f \text{ is convex, } L_{\|\cdot\|_p}(f) \leq L\}$$

Assume that when solving  $(P_f)$ ,  $f \in \mathcal{F}_{n,p}(L)$  is learned via calls, one per step, to a FO (or any local) oracle. Then for every  $t \leq n$  and any  $t$ -step algorithm  $\mathcal{B}$  one has

$$\sup_{f \in \mathcal{F}_{n,p}(L)} [f(x_{\mathcal{B}}(f)) - \text{Opt}(f)] \geq 0.01LR/t^{1/p}$$

- $x_{\mathcal{B}}(f)$ : solution generated in  $t$  steps by  $\mathcal{B}$  as applied to  $(P_f)$

$\Rightarrow$  As  $p > 2$  grows, our abilities to minimize oracle-represented nonsmooth convex functions over  $\|\cdot\|_p$ -balls *at a dimension independent rate* deteriorate and disappear at  $p = \infty$ .

# Favorable Geometry: Illustration

♣ The most attractive feature of MD is *ability to adjust itself, to some extent, to problem's geometry and to ensure, under favorable circumstances, (nearly) dimension independent rate of convergence*. For example:

- When minimizing convex  $f$  over  $\ell_2$ -ball  $\{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$ , MD with *Euclidean setup* ensures

$$f(x^t) - \min_{x \in X} f(x) \leq O(1) [\max_X f - \min_X f] / \sqrt{t}$$

- When minimizing convex  $f$  over  $\ell_1$ -ball  $\{x \in \mathbf{R}^n : \|x\|_1 \leq 1\}$ , MD with appropriate *Non-Euclidean setup* ensures

$f(x^t) - \min_{x \in X} f(x) \leq O(1) \sqrt{\ln(n+1)} [\max_X f - \min_X f] / \sqrt{t}$ ,  
and similarly for minimizing over nuclear norm ball in  $\mathbf{R}^{n \times n}$ .

- “Wrong setup” (Euclidean when minimizing over  $\ell_1$ /nuclear norm ball, or  $\ell_1$ /nuclear norm when minimizing over  $\ell_2$ -ball) can spoil the efficiency by factor as large as  $O(\sqrt{n/\ln(n)})$ .

# Stochastic case

♣ **Situation:** Given  $X \subset E$  and proximal setup  $\|\cdot\|, \omega(\cdot)$ , we want to process vector fields  $g_\tau(x) : X \rightarrow E$  represented by *Stochastic Oracle*. At  $\tau$ -th call to SO, the query point being  $x_\tau \in X$ , the oracle returns *an estimate*  $h_\tau(x_\tau; \xi_\tau) \in E$  of  $g_\tau(x_\tau)$ . Here  $h_\tau(\cdot; \cdot)$  are deterministic functions, and  $\xi_1, \xi_2, \dots$  are *i.i.d.* disturbances.

♠ **Example:** Problem  $\min_{x \in X} [f(x) = \mathbf{E}_{\xi \sim P} F(x, \xi)]$  with convex in  $x \in X$  integrant  $F$ .

The associated vector field  $g(x) = f'(x)$  is usually difficult to compute. However, assuming one can sample from  $P$  and  $F$  is easy to compute, we can set

$$h_\tau(x; \xi_\tau) = F'_x(x, \xi_\tau) \text{ with } \xi_1, \xi_2, \dots \text{ drawn from } P$$

♠ **Standing Assumption:** When processing  $\{g_\tau(\cdot)\}_\tau$ , for some  $L, \sigma, \mu$  and all  $x \in X, \tau$  it holds:

$$\|g_\tau(x)\|_* \leq L, \|\mathbf{E}_\xi \{\Delta_\tau(x; \xi)\}\|_* \leq \mu, \mathbf{E}_\xi \{\|\Delta_\tau(x; \xi)\|_*^2\} \leq \sigma^2$$

- $\Delta_\tau(x; \xi) := h_\tau(x; \xi) - g_\tau(x)$ : oracle's error

# Stochastic Mirror Descent

- $X$ : convex compact subset of Euclidean space  $E$
- $\|\cdot\|, \omega(\cdot)$ : proximal setup for  $(E, X) \Rightarrow \Omega = \sqrt{2[\max_X \omega - \min_X \omega]}$
- $\{g_\tau(x) : X \rightarrow E\}_\tau$ : vector fields of interest,  $\|g_\tau(x)\|_* \leq L < \infty$
- $\{h_\tau(x; \xi) = g_\tau(x) + \Delta_\tau(x; \xi) : X \times \Xi \rightarrow E\}_\tau$ : Stochastic oracle  
 $\|\mathbf{E}_{\xi \sim P} \Delta_\tau(x; \xi)\|_* \leq \mu, \mathbf{E}_{\xi \sim P} \{\|\Delta_\tau(x; \xi)\|_*^2\} \leq \sigma^2$

♣ **Stochastic Mirror Descent** is the recurrence

$$x_1 = x_\omega := \operatorname{argmin}_X \omega; x_{\tau+1} = \operatorname{Prox}_{x_\tau}(\gamma_\tau h_\tau(x_\tau; \xi_\tau))$$
$$x^t = \sum_{\tau \leq t} \lambda_\tau^t x_\tau, \lambda_\tau^t = \gamma_\tau / \sum_{s \leq t} \gamma_s$$

- $\xi_\tau \sim P$ : independent
- $\gamma_\tau > 0$ : deterministic stepsizes

# Stochastic Mirror Descent (continued)

$$x_1 = x_\omega := \operatorname{argmin}_X \omega; x_{\tau+1} = \operatorname{Prox}_{x_\tau}(\gamma_\tau [g_\tau(x_\tau) + \Delta_\tau(x_\tau; \xi_\tau)])$$

$$x^t = \sum_{\tau \leq t} \lambda_\tau^t x_\tau, \quad \lambda_\tau^t = \gamma_\tau / \sum_{s \leq t} \gamma_s$$

$$\|g_\tau(x)\|_* \leq L, \quad \|\mathbf{E}_{\xi \sim P} \Delta_\tau(x; \xi)\|_* \leq \mu, \quad \mathbf{E}_{\xi \sim P} \{\|\Delta_\tau(x; \xi)\|_*^2\} \leq \sigma^2$$

## ♣ Main Property of SMD: *One has*

$$\mathbf{E} \left\{ \mathcal{E}_t := \max_{z \in X} \sum_{\tau \leq t} \lambda_\tau^t \langle g(x_\tau), x_\tau - z \rangle \right\} \\ \leq \frac{\Omega^2 + [L^2 + 2\sigma^2] \sum_{\tau \leq t} \gamma_\tau^2}{\sum_{\tau \leq t} \gamma_\tau} + 2\mu\Omega$$

- *In particular,  $\gamma_\tau = \Omega / \sqrt{[L^2 + 2\sigma^2]t}$ ,  $1 \leq \tau \leq t$ , yields*

$$\mathbf{E}\{\mathcal{E}_t\} \leq \Theta / \sqrt{t} + 2\mu\Omega, \quad \Theta = 2\Omega\sqrt{L^2 + 2\sigma^2}.$$

- *Strengthening the bound on the second moment of  $\|\Delta_\tau\|_*$  to  $\mathbf{E}\{\exp\{\|\Delta_\tau\|_*^2/\sigma^2\}\} \leq \exp\{1\}$ , large deviation probabilities obey an exponential bound:*

$$\forall \theta > 0 : \operatorname{Prob} \left\{ \mathcal{E}_t > [\Theta + \theta\Sigma] / \sqrt{t} + 2\mu\Omega \right\} \leq O(1)e^{-\theta} \\ [\Sigma = 4\Omega\sigma]$$

# Stochastic Mirror Descent (continued)

♣ When  $g_\tau(\cdot) \equiv g(\cdot)$  is associated with a problem with convex structure, e.g.,

A.  $\min_{x \in X} f(x) \Rightarrow g(x) = f'(x)$ , or

B.  $\min_{u \in U} \max_{v \in V} f(u, v) \Rightarrow g(u, v) = [f'_u(u, v); -f'_v(u, v)]$ ,

the residual  $\mathcal{E}_t$  upper-bounds inaccuracy of the approximate solution  $x^t$  to the problem of interest.

$\Rightarrow$  *t*-step SMD allows to solve stochastic convex problems with *expected inaccuracy*  $O(1/\sqrt{t})$ . For example,

- in the case of A, we get

$$\mathbf{E}\{f(x^t) - \min_X f\} \leq 2\Omega\sqrt{L^2 + 2\sigma^2}/\sqrt{t} + 2\mu\Omega$$

- in the case of B, we get

$$\mathbf{E}\{[\bar{f}(u^t) - \min_U \bar{f}] + [\max_V \underline{f} - \underline{f}(v^t)]\} \leq 2\Omega\sqrt{L^2 + 2\sigma^2}/\sqrt{t} + 2\mu\Omega.$$

♠ **Note:** In typical stochastic problems, in every dimension, not only a large one,  $O(1/\sqrt{t})$  is the *best rate allowed by Statistics*.

# Stochastic Mirror Descent: Illustration

♣ Consider *Binary Classification problem* where we draw from a distribution  $P$  examples  $\xi_\tau = (\eta_\tau, y_\tau) \in \mathbf{R}^N \times \{\pm 1\}$  and want to build a *linear classifier*  $y \sim \text{sign}(\langle x, \eta \rangle)$ .

♠ The problem can be modeled as

$$\text{Opt}(\rho) = \min_{\|x\| \leq 1} [\rho_\rho(x) = \rho(\rho x) := \mathbf{E}\{\max[1 - y\langle \rho x, \eta \rangle, 0]\}]$$

$[\rho(x) : \text{convex upper bound on the probability for } x \text{ to mis-classify}]$

• Let  $\|\cdot\|$  be (a)  $\|\cdot\|_2$ , or (b)  $\|\cdot\|_1$ , or (c) nuclear norm on  $\mathbf{R}^N = \mathbf{R}^{m \times n}$

♠ Assuming  $\mathbf{E}\{\|\eta\|_*^2\} \leq R^2 < \infty$  and setting

$$h(x; \eta, y) \equiv -\rho y \chi(1 - y\langle \rho x, \eta \rangle > 0) \eta,$$

$$g(x) := \mathbf{E}_{\eta, y}\{h(x; \eta, y)\} \in \rho'_\rho(x)$$

we satisfy Standing Assumption with

$$X = \{\|x\| \leq 1\}, \quad L = \rho R, \quad \sigma = 2\rho R, \quad \mu = 0.$$

$\Rightarrow$  For every  $t \geq 1$ , drawing a  $t$ -element sample from  $P$  and applying  $t$ -step SMD with appropriate proximal setup, we get a linear classifier  $\rho x^t$ ,  $\|x^t\| \leq 1$ , such that

$$\mathbf{E}\{\rho(\rho x^t)\} \leq \text{Opt}(\rho) + \rho R t^{-1/2} \times \begin{cases} O(1), & \text{case (a)} \\ O(1)\sqrt{\ln(N)}, & \text{cases (b), (c)} \end{cases}$$



# Utilizing Problem's Structure: Mirror Prox

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

♣ Unimprovable or not, convergence rate  $O(1/\sqrt{t})$  is slow. When we can do better?

- One can use *bundle* versions of MD re-utilizing past information. In practice, this improves the convergence pattern at the price of *controlled* increase in the computational cost of a step. *Theoretical complexity bounds, however, remain intact.*

- When  $f$  is smooth:  $\|f'(x) - f'(x')\|_* \leq \mathcal{M}\|x - x'\|$ , the MD efficiency improves to  $f(x^t) - \min_X f \leq \Omega^2 \mathcal{M}/t$ . *This is of no actual interest: with Nesterov's optimal method for smooth convex minimization one achieves unimprovable in the large-scale case efficiency  $O(1)\Omega^2 \mathcal{M}/t^2$ .*

- When  $f$  is *strongly convex*, properly modified MD converges at the rate  $O(1/t)$ .

- *For a wide spectrum of "well-structured"  $f$ , rate  $O(1/t)$  can be achieved by *smooth* saddle point reformulation of (P).*

♣ **Situation:**  $X$  is a convex compact subset of Euclidean space  $E$ ,  $\|\cdot\|, \omega(\cdot)$  is a proximal setup,  $g(\cdot) : X \rightarrow E$  is a vector field represented by an oracle.

• At  $\tau$ -th call,  $x_\tau \in X$  being the query point, the oracle returns an estimate  $h(x_\tau; \xi_\tau) = g(x_\tau) + \Delta(x_\tau; \xi_\tau)$  of  $g(x_\tau)$ ,  $\xi_\tau$  are i.i.d.,

$$\|\mathbf{E}_\xi\{\Delta(x; \xi)\}\|_* \leq \mu, \quad \mathbf{E}_\xi\{\|\Delta(x; \xi)\|_*^2\} \leq \sigma^2, \quad \forall x \in X$$

•  $g(\cdot)$  satisfies

$$\|g(x) - g(x')\|_* \leq \mathcal{M}\|x - x'\| + L \quad \forall (x, x' \in X)$$

• **Note:**  $L = \sigma = \mu = 0 \Leftrightarrow g(\cdot)$  is *Lipschitz & precisely* observed.

♣ **Mirror Prox** is the recurrence

$$x_1 = x_\omega;$$

$$x_\tau \mapsto w_\tau = \text{Prox}_{x_\tau}(\gamma_\tau h(x_\tau; \xi_{2\tau-1}))$$

$$\mapsto x_{\tau+1} = \text{Prox}_{x_\tau}(\gamma_\tau h(w_\tau; \xi_{2\tau}))$$

$$x^t = \sum_{\tau \leq t} \lambda_\tau^t w_\tau, \quad \lambda_\tau^t = \gamma_\tau / \sum_{s \leq t} \gamma_s$$

with deterministic stepsizes  $\gamma_\tau > 0$ .

# Mirror Prox (continued)

- $X \subset E, \|\cdot\|, \omega \Rightarrow \Omega$
- $g(\cdot) : X \rightarrow E: \|g(x) - g(x')\|_* \leq \mathcal{M}\|x - x'\| + L$
- oracle  $x \mapsto h(x; \xi) = g(x) + \Delta(x; \xi)$ :  
 $\|\mathbf{E}_\xi\{\Delta(x; \xi)\}\|_* \leq \mu, \mathbf{E}_\xi\{\|\Delta(x; \xi)\|_*^2\} \leq \sigma^2$
- $x_\tau \mapsto w_\tau = \text{Prox}_{x_\tau}(\gamma_\tau h(x_\tau; \xi_{2\tau-1})) \mapsto x_{\tau+1} = \text{Prox}_{x_\tau}(\gamma_\tau h(w_\tau; \xi_{2\tau}))$   
 $x^t = \sum_{\tau \leq t} \lambda_\tau^t w_\tau, \lambda_\tau^t = \gamma_\tau / \sum_{s \leq t} \gamma_s$

♣ **Main Property of MP:** Let  $0 < \gamma_\tau \leq \frac{1}{2\mathcal{M}}$ . Then

$$\mathbf{E} \left\{ \mathcal{E}_t := \max_{z \in X} \sum_{\tau \leq t} \lambda_\tau^t \langle g(x_\tau), x_\tau - z \rangle \right\} \leq \frac{\Omega^2 + [3L^2 + 7\sigma^2] \sum_{\tau \leq t} \gamma_\tau^2}{\sum_{\tau \leq t} \gamma_\tau} + 2\mu\Omega$$

• In particular,  $\gamma_\tau = \min \left[ (2\mathcal{M})^{-1}, \Omega / \sqrt{[3L^2 + 7\sigma^2]t} \right], \tau \leq t$ , yields

$$\mathbf{E}\{\mathcal{E}_t\} \leq 2\Omega^2\mathcal{M}/t + \Theta/\sqrt{t} + 2\mu\Omega, \quad \Theta = 2\Omega\sqrt{3L^2 + 7\sigma^2}.$$

♠ **Note:** In the smooth deterministic case  $L = \sigma = \mu = 0$ , we get  $O(1/t)$  convergence!

# Mirror Prox (continued)

- $X \subset E, \|\cdot\|, \omega \Rightarrow \Omega$
- $g(\cdot) : X \rightarrow E: \|g(x) - g(x')\|_* \leq \mathcal{M}\|x - x'\| + L$
- oracle  $x \mapsto h(x; \xi) = g(x) + \Delta(x; \xi)$ :  
$$\|\mathbf{E}_\xi\{\Delta(x; \xi)\}\|_* \leq \mu, \quad \mathbf{E}_\xi\{\|\Delta(x; \xi)\|_*^2\} \leq \sigma^2$$
- $x_\tau \mapsto w_\tau = \text{Prox}_{x_\tau}(\gamma_\tau h(x_\tau; \xi_{2\tau-1})) \mapsto x_{\tau+1} = \text{Prox}_{x_\tau}(\gamma_\tau h(w_\tau; \xi_{2\tau}))$   
$$x^t = \sum_{\tau \leq t} \lambda_\tau^t w_\tau, \quad \lambda_\tau^t = \gamma_\tau / \sum_{s \leq t} \gamma_s$$

♠  $\gamma_\tau = \min \left[ (2\mathcal{M})^{-1}, \Omega / \sqrt{[3L^2 + 7\sigma^2]t} \right], 1 \leq \tau \leq t$ , yields

$$\mathbf{E}\{\mathcal{E}_t\} \leq \frac{2\Omega^2\mathcal{M}}{t} + \frac{\Theta}{\sqrt{t}} + 2\mu\Omega, \quad \Theta = 2\Omega\sqrt{3L^2 + 7\sigma^2}.$$

• Strengthening the bound on the second moment of  $\|\Delta\|_*$  to  $\mathbf{E}\{\exp\{\|\Delta\|_*^2/\sigma^2\}\} \leq \exp\{1\}$ , large deviation probabilities obey an exponential bound:

$$\forall \theta > 0 : \text{Prob} \left\{ \mathcal{E}_t > 2\Omega^2\mathcal{M}/t + [\Theta + \theta\Sigma]/\sqrt{t} + 2\mu\Omega \right\} \leq O(1)e^{-\theta}$$
$$[\Sigma = 9\Omega\sigma]$$

# Application: $O(1/t)$ Nonsmooth Convex Minimization

$$\text{Opt}(P) = \min_{u \in U} f(u) \quad (P)$$

♣ **Corollary:** Let  $(P)$  be a convex program with compact  $U \subset E_U$  and with  $f$  such that

$$f(u) = \max_{v \in V} \phi(u, v)$$

- $V$ : compact convex subset of Euclidean space  $E_V$
- $\phi(u, v)$ : convex-concave with *Lipschitz continuous* gradient so that  $(P)$  is the primal form of the saddle point problem

$$\min_{u \in U} \max_{v \in V} \phi(u, v) \quad (SP)$$

The vector field  $g(u, v) = [\phi'_u(u, v); -\phi'_v(u, v)]$  associated with  $(SP)$  is Lipschitz continuous. Equipping

- $E := E_U \times E_V, X := U \times V$  — with a proximal setup  $\|\cdot\|, \omega$ ,
  - $g(\cdot)$  — with a precise deterministic oracle,
- $t$ -step MP yields  $(u^t, v^t) \in U \times V$  such that

$$f(u^t) - \text{Opt}(P) \leq O(1)\Omega\mathcal{M}/t$$

$$\mathcal{M} = \min\{M : \|g(x) - g(x')\|_* \leq M\|x - x'\| \ \forall (x, x' \in X)\}$$

$$\min_{u \in U} [f(u) = \max_{v \in V} \phi(u, v)]$$

♣ **Fact:** If  $\phi(u, v)$  is

- convex-concave with Lipschitz continuous gradient,
- affine in  $u$ ,
- strongly concave in  $v$ ,

then properly modified MP ensures  $O(1/t^2)$  convergence rate.

♠ **Note:** The premise does *not* imply smoothness of  $f$ .

# Smooth and Bilinear Saddle Point Representations

♣ **Fact:** Representations  $f(u) = \max_{v \in V} \phi(u, v)$  with smooth convex-concave, and even with bilinear  $\phi$  are available for wide spectrum of convex functions  $f$ . Whenever it is the case,  $f$  can be minimized via MP at the rate  $O(1/t)$ .

•  $f(u) = \max_{k \leq K} f_k(u)$  with smooth convex  $f_k$

$$\Rightarrow f(u) = \max_{v \geq 0, \sum_k v_k = 1} \sum_k v_k f_k(u)$$

•  $f(u) = \|Au - b\| \Rightarrow f(u) = \max_{\|v\|_* \leq 1} \langle v, Au - b \rangle$

•  $f(u) = \|u\| + \frac{1}{2} \|Au - b\|_2^2$

$$\Rightarrow f(u) = \max_{\|v\|_* \leq 1, w} [\langle u, v \rangle + \langle w, Au - b \rangle - \frac{1}{2} w^T w]$$

•  $f(u)$ : sum of  $k$  largest eigenvalues of  $\mathcal{A}(u) = Au - b \in \mathbf{S}^n$

$$\Rightarrow f(u) = \max_v [\text{Tr}(v \mathcal{A}(u)) : 0 \preceq v \preceq I_n, \text{Tr}(v) = k]$$

•  $f(u) = \inf_{b \in \mathbf{R}} \left[ \frac{1}{N} \sum_{i=1}^N \max[1 - y_i(\langle u, \eta_i \rangle + b), 0] \right]$

$$\Rightarrow f(u) = \max_{v \in V} \sum_{i=1}^N v_i [1 - y_i \langle u, \eta_i \rangle]$$

$$V = \{v : 0 \leq v_i \leq 1/N \forall i, \sum_i y_i v_i = 0\} \subset \{v \in \mathbf{R}^N : \|v\|_1 \leq 1\}$$

# $O(1/t)$ Nonsmooth Convex Minimization: Comments

$$\text{Opt}(P) = \min_{u \in U} f(u) \quad (P)$$

- Convex programs always have a lot of structure (otherwise, how could we know that the problem is convex?)

*Accelerating algorithms by utilizing problem's structure is an old and still challenging goal.*

- A common way to utilize structure is via “structure-revealing” *conic* formulations (Linear/Conic Quadratic/Semidefinite) and Interior Point Methods. However, *in the large scale case IPM iteration may become prohibitively costly.*

- Utilizing structure *within the realm of oracle-oriented methods with computationally cheap iterations* is due to Nesterov (2003).

Nesterov's Smoothing (2003) uses saddle point representation of a nonsmooth  $f$  to approximate  $f$  by a *smooth* function which is further minimized by Nesterov's algorithm for smooth convex minimization. The resulting convergence rate is  $O(1/t)$ .

- MP offers another way to utilize saddle point representation to achieve the same  $O(1/t)$  rate.

“Practical scopes” of these two approaches are nearly identical.



# $O(1/t)$ Nonsmooth Convex Minimization: Examples

## ♣ Problem of interest:

$$\text{Opt}(P) = \min_{\|u\| \leq 1} \|Au - b\|_p, \quad A : M \times N \quad (P)$$

where  $p = 2$  or  $p = \infty$ , and  $\|\cdot\|$  is

(a)  $\|\cdot\|_2$  on  $\mathbf{R}^N$ , or (b)  $\|\cdot\|_1$  on  $\mathbf{R}^N$ , or (c) nuclear norm on  $\mathbf{R}^N = \mathbf{R}^{m \times n}$

## ♠ Bilinear saddle point reformulation is

$$\text{SadVal} = \min_{u \in U} \max_{v \in V} \langle v, Au - b \rangle$$

$$U = \{\|u\| \leq 1\}, \quad V = \{\|v\|_q \leq 1\}, \quad q = \frac{p}{p-1} \in \{1, 2\}$$

and its domain is the product of two favorable geometry atoms.

♠ Applying  $t$ -step MP with appropriate setup, we get  $u^t$  with  $\|u^t\| \leq 1$  and

$$f(u^t) - \text{Opt}(P) \leq \kappa \|A\|_{\|\cdot\|, p} / t$$

$$\|A\|_{\|\cdot\|, p} = \max\{\|Au\|_p : \|u\| \leq 1\}$$

$$\kappa = O(1) \ln^{1/2-1/p}(M+1) \times \begin{cases} 1, & \text{case (a)} \\ \sqrt{\ln(N+1)}, & \text{case (b)} \\ \sqrt{\ln(m+1)}, & \text{case (c)} \end{cases}$$

# $O(1/t)$ Nonsmooth Convex Minimization: Examples

$$\text{Opt}(P) = \min_{\|u\| \leq 1} \|Au - b\|_p, \quad A : M \times N, \quad p \in \{2, \infty\} \quad (P)$$

$\|\cdot\|$ : (a)  $\|\cdot\|_2$  on  $\mathbf{R}^N$  | (b)  $\|\cdot\|_1$  on  $\mathbf{R}^N$  | (c) nuclear norm on  $\mathbf{R}^N = \mathbf{R}^{m \times n}$

$$\Rightarrow f(u^t) - \text{Opt}(P) \leq O(1) \ln(MN) \|A\|_{\|\cdot\|, p} / t$$

♠ MP step reduces to computing  $O(1)$  matrix-vector products involving  $A$  and  $A^*$ , plus

—  $O(M + N)$  a.o. in cases (a), (b)

— computing svd's of two  $m \times n$  matrices in case (c).

⇒ Except for case (c), *MP is computationally cheap...*

♠ **Note:** When solving a *Least Squares* problem

$$(LS) \quad \text{Opt}(A, b) = \min_{\|u\|_2 \leq 1} \|Au - b\|_2 \quad [A : n \times n]$$

with  $A$  represented by *multiplication oracle*  $u, u' \mapsto Au, A^T u'$ ,

the rate  $O(1/t)$  is unimprovable in the large-scale case:

• *The worst-case, over  $(A, b)$  with  $\|A\|_{2,2} \leq 1$  and  $\text{Opt}(A, b) = 0$ , inaccuracy in terms of the objective of (LS) is, for every  $t$ -step algorithm, at least  $O(1)/t$ , provided  $t \leq n/4$ .*

# Acceleration by Randomization

## ♣ Problem of interest:

$$\text{Opt} = \min_{\|u\|_1 \leq 1} \|Au - b\|_p \quad [A : m \times n, p \in \{2, \infty\}]$$

$$\Leftrightarrow (\ell_1) : \min_{\|u\|_1 \leq 1} \max_{\|v\|_{p/(p-1)} \leq 1} \langle v, Au - b \rangle$$

$$\Rightarrow g(u, v) = [A^T v; b - Au] : X := U \times V \rightarrow \mathbf{R}^{m+n}$$

$$U = \{u : \|u\|_1 \leq 1\}, V = \{v : \|v\|_{p/(p-1)} \leq 1\}.$$

♠ *Omitting from now on logarithmic in  $m, n$  factors, MP solves  $(\ell_1)$  within accuracy  $\varepsilon$  in*

$$N(\varepsilon) = \|A\|_{1,p}/\varepsilon, \quad \|A\|_{1,p} = \max_{j \leq n} \|\text{Col}_j[A]\|_p$$

*steps*, with two multiplications of vectors from  $U$  and from  $V$  by  $A, A^T$ , plus  $O(m+n)$  a.o. “overhead,” per step.

$\Rightarrow$  *The arithmetic cost of  $\varepsilon$ -solution for a general-type  $A$  is*

$$C_d(\varepsilon) = mn \|A\|_{1,p} / \varepsilon \text{ a.o.}$$

In fact, this is the best operation count achievable in the large-scale case with known so far *deterministic* algorithms.

• *For large  $m, n$ , matrix-vector multiplications may become too time consuming...*

# Acceleration by Randomization (continued)

♠ Matrix-vector multiplications are easy to randomize:

*In order to compute  $Bu$ ,  $B \in \mathbf{R}^{M \times N}$ , we draw an index  $j$  at random according to*

$$\text{Prob}\{j = j\} = u_j / \|u\|_1, \quad 1 \leq j \leq N$$

*and return the vector*

$$h = \|u\|_1 \text{sign}(u_j) \text{Col}_j[B]$$

**Note:**

- $\mathbf{E}\{h\} = Bu$ ,  $\|h\|_q \leq \|u\|_1 \|B\|_{1,q}$
- Generating  $h$  costs  $O(1)(M + N)$  a.o. (assuming cost  $O(1)$  of computing/extracting individual entry of  $B$ ).

# Acceleration by Randomization (continued)

$$\begin{aligned} \text{Opt} &= \min_{\|u\|_1 \leq 1} \|Au - b\|_p \quad [A : m \times n, p \in \{2, \infty\}] \\ \Leftrightarrow (\ell_1) : \quad &\min_{\|u\|_1 \leq 1} \max_{\|v\|_{p/(p-1)} \leq 1} \langle v, Au - b \rangle \\ \Rightarrow \quad &g(u, v) = [A^T v; b - Au] : X := U \times V \rightarrow \mathbf{R}^{m+n} \\ &U = \{u : \|u\|_1 \leq 1\}, V = \{v : \|v\|_{p/(p-1)} \leq 1\}. \end{aligned}$$

♠ When solving  $(\ell_1)$  with  $p = \infty$  by MP with the precise values of  $g(\cdot)$  replaced with their cheap unbiased random estimates, we  $(1 - \delta)$ -reliably get  $\varepsilon$ -solution to  $(\ell_1)$  in  $\ln(1/\delta) [\|A\|_{1,\infty}/\varepsilon]^2$  steps, the total computational effort being

$$C_r = (m + n) \ln(1/\delta) [\|A\|_{1,\infty}/\varepsilon]^2 \text{ a.o.}$$

- ♠ The “deterministic” operation count is  $C_d = mn\|A\|_{1,\infty}/\varepsilon$ .
- $\Rightarrow$  With the relative accuracy  $\varepsilon/\|A\|_{1,\infty}$  and  $\delta$  fixed and  $m, n$  large, randomized algorithm by far outperforms its deterministic competitors.
- In addition, Randomized MP exhibits **sublinear time behavior**: when  $m, n$  are large,  $\varepsilon$ -solution is obtained, in a  $(1 - \delta)$ -reliable fashion, by inspecting **negligibly small** fraction of the  $mn$  data entries.

# Acceleration by Randomization (continued)

♠ In the case of  $p = \infty$ , our construction basically recovers the *ad hoc* sublinear time algorithm for matrix games (Grigoriadis & Khachiyan, 1994).

♠ *In the case of  $p = 2$* , randomization leads to iteration count  $\ln(1/\delta)[\|A\|_{1,2}/\varepsilon]^2 \Gamma^2[A]$ ,  $\Gamma(A) = \sqrt{m}\|A\|_{1,\infty}/\|A\|_{1,2} \in [1, \sqrt{m}]$  and operation count  $\mathcal{C}_r = (m+n) \ln(1/\delta)[\|A\|_{1,2}/\varepsilon]^2 \Gamma^2[A]$  a.o. vs. the “deterministic” operation count  $\mathcal{C}_d = mn[\|A\|_{1,2}/\varepsilon]$  a.o.

- with  $\Gamma[A]$  like  $O(1) \ln(mn)$ , everything is as when  $p = \infty$
- with  $\Gamma[A]$  as large as  $O(\sqrt{m})$ , randomization is really bad.

♠ **However:** Preprocessing  $[A, b] \Rightarrow [\bar{A}, \bar{b}] = \mathbf{F} \text{Diag}\{\chi\}[A, b]$  with  $m \times m$  DFT matrix  $\mathbf{F}$  and  $\chi \sim \text{Uniform}(\{-1; 1\}^m)$  yields *equivalent* problem and ensures  $(1 - \delta)$ -reliably  $\Gamma[\bar{A}] \leq \sqrt{\ln(mn/\delta)}$ .

$\Rightarrow$  *With randomization and preprocessing, the operation count is*

$$\mathcal{C}_r = mn + (m+n) \ln^2(1/\delta)[\|A\|_{1,2}/\varepsilon]^2$$

*which for small and fixed  $\varepsilon/\|A\|_{1,2}$  and large  $m, n$  is negligibly small as compared to  $\mathcal{C}_d = mn[\|A\|_{1,2}/\varepsilon]$  a.o.*

# How it Works: 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

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

Burglar wants to maximize his expected profit

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

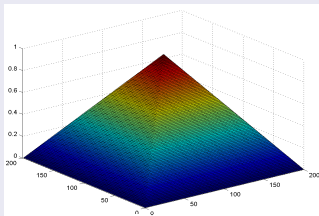
the interest of Policeman is completely opposite.

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

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

$$\min_{\substack{u \geq 0, \\ \sum_{j=1}^n u_j = 1}} \max_{\substack{v \geq 0, \\ \sum_{i=1}^n v_i = 1}} \phi(u, v) := v^T A u$$
$$A_{ij} = w_i(1 - \exp\{-\theta \text{dist}(i, j)\})$$

# Policeman vs. Burglar (continued)



Wealth on  $n \times n$  square grid of houses

	IPM	MP	Rand MP
$N$	Steps/CPU, sec/ $\epsilon$	Steps/CPU, sec/ $\epsilon$	Steps/CPU, sec/ $\epsilon$
1600	21/120/6.0e-9	78/6/1.0e-3	10556/264/1.0e-3
6400	21/6930/1.1e-8	80/31/1.0e-3	10408/796/1.0e-3
14400	not tested	95/171/1.0e-3	9422/1584/1.0e-3
40000	out of memory	15 <sup>†</sup> /5533 <sup>†</sup> /0.022 <sup>†</sup>	10216/4931/1.0e-3

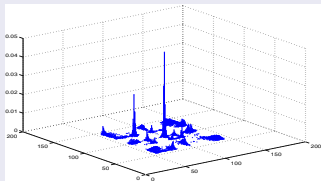
## Policeman vs. Burglar, $N$ houses

Target residual  $\epsilon_t \leq 1.e-3$  IPM: mosekopt

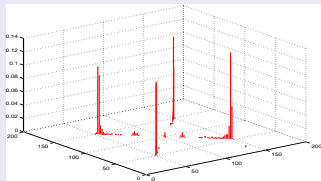
<sup>†</sup>: termination when reaching the CPU limit of 5,400 sec



# Policeman vs. Burglar (continued)

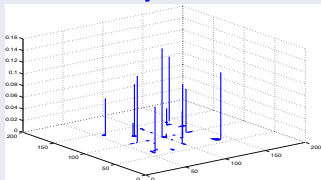


**Policeman**

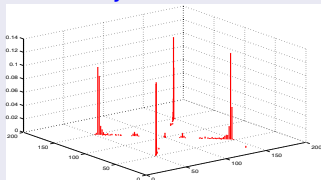


**Burglar**

🔥 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**

**200 × 200 grid of houses**

# References

- A. Beck, M. Teboulle, Mirror Descent... – *OR Letters* **31** '03
- M. Grigoriadis, L. Khachiyan, A Sublinear Time... – *OR Letters* **18** '95
- A. Juditsky, F. Kılınç Karzan, A. Nemirovski,  $\ell_1$  Minimization... ('11), <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, 2012
- A. Nemirovski, Efficient Methods... – *Ekonomika i Mat. Metody* **15** '79
- A. Nemirovski, D. Yudin, Problem Complexity... – J. Wiley '83
- 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

**Tutorial: Mirror Descent Algorithms for Large-Scale Deterministic and Stochastic Convex Optimization**  
**Selected Proofs**

## 1 Mirror Descent: Stochastic Case

**Situation:** At  $t$ -th call to Stochastic Oracle, the query point being  $x_t \in X$ , the SO returns vector  $g_t(x_t) + \Delta_t(x_t, \xi_t)$ , with independent  $\xi_t \sim P$ ,  $t = 1, 2, \dots$ . Besides this, for all  $x \in X$  and all  $t$  it holds

$$\|g_t(\cdot)\|_* \leq L < \infty, \mathbf{E}_\xi\{\|\Delta_t(x, \xi)\|_*^2\} \leq \sigma^2, \|\mathbf{E}_\xi\{\Delta_t(x, \xi)\}\|_* \leq \mu. \quad (1)$$

For the MD recurrence

$$x_1 = x_\omega; x_{t+1} = \text{Prox}_{x_t}(\gamma_t[g_t(x_t) + \Delta_t(x_t, \xi_t)]) \quad (2)$$

with deterministic stepsizes  $\gamma_t > 0$ , we have, setting  $g_t = g_t(x_t)$ ,  $\Delta_t = \Delta_t(x_t, \xi_t)$ :

$$\begin{aligned} & \gamma_t \langle g_t + \Delta_t, x_{t+1} - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) - V_{x_t}(x_{t+1}) \text{ [see (2) in Transparencies]} \\ \Rightarrow & \gamma_t \langle g_t + \Delta_t, x_t - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) + [\gamma_t \langle g_t + \Delta_t, x_t - x_{t+1} \rangle - V_{x_t}(x_{t+1})] \\ \Rightarrow & \gamma_t \langle g_t + \Delta_t, x_t - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) + \frac{1}{2} \gamma_t^2 \|g_t + \Delta_t\|_*^2 \\ \Rightarrow & \sum_{t=1}^T \gamma_t \langle g_t, x_t - x \rangle \leq \frac{1}{2} \Omega^2 + \frac{1}{2} \sum_{t=1}^T \|g_t + \Delta_t\|_*^2 + \sum_{t=1}^T \gamma_t \langle \Delta_t, x - x_t \rangle \\ \Rightarrow & \left[ \sum_{t=1}^T \gamma_t \right] \epsilon_T \leq \frac{1}{2} \Omega^2 + \frac{1}{2} \sum_{t=1}^T \|g_t + \Delta_t\|_*^2 + \max_{x \in X} \sum_{t=1}^T \gamma_t \langle \Delta_t, x - x_t \rangle \\ & \epsilon_T := \max_{x \in X} \sum_{t=1}^T \lambda_t \langle g_t, x_t - x \rangle, \lambda_t = \gamma_t / \sum_{s=1}^T \gamma_s. \end{aligned}$$

The bottom line is that

$$\left[ \sum_{t=1}^T \gamma_t \right] \epsilon_T \leq \frac{1}{2} \Omega^2 + \frac{1}{2} \sum_{t=1}^T \|g_t + \Delta_t\|_*^2 + \max_{x \in X} \sum_{t=1}^T \gamma_t \langle \Delta_t, x - x_t \rangle \quad (3)$$

**Our goal** is to prove the following

**Theorem 1.1** (i) Assuming (1), for the recurrence (2) for every  $T = 1, 2, \dots$  one has

$$\mathbf{E}\{\epsilon_T\} \leq \Theta := \frac{\Omega^2 + [L^2 + \frac{3}{2}\sigma^2] \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t} + 2\mu\Omega. \quad (4)$$

(ii) Strengthening (1) to

$$\|g_t(\cdot)\|_* \leq L < \infty, \mathbf{E}_\xi\{\exp\{\|\Delta_t(x, \xi)\|_*^2/\sigma^2\}\} \leq \exp\{1\}, \|\mathbf{E}_\xi\{\Delta_t(x, \xi)\}\|_* \leq \mu \quad (5)$$

for all  $x \in X$  and all  $t$ , we have for every  $\theta > 0$ :

$$\begin{aligned} \text{Prob}\{\epsilon_T > \Xi + \theta\Upsilon\} &\leq 6 \exp\{-\theta\} + \exp\{-\theta^2/4\}, \\ \Upsilon &= 2 \frac{\sigma^2 \sum_{t=1}^T \gamma_t^2 + \sigma\Omega \sqrt{\sum_{t=1}^T \gamma_t^2}}{\sum_{t=1}^T \gamma_t}. \end{aligned} \quad (6)$$

Note that with the stepsizes

$$\gamma_t = \frac{\Omega}{\sqrt{L^2 + 3\sigma^2/2}\sqrt{T}}, \quad 1 \leq t \leq T \quad (7)$$

one has

$$\Xi = \frac{2\Omega\sqrt{L^2 + 3\sigma^2/2}}{\sqrt{T}} + 2\mu\Omega, \quad \Upsilon \leq 4 \frac{\Omega\sigma}{\sqrt{T}}. \quad (8)$$

**Proof.**

1<sup>0</sup>. We need the following

**Lemma 1.1** Given deterministic  $\gamma_t$  and (perhaps, stochastic)  $g_t, \Delta_t$  such that  $\|g_t\|_* \leq L < \infty$  and

$$\mathbf{E}\{\|\Delta_t\|_*^2\} \leq \sigma^2 \forall t, \quad (9)$$

one has

$$\mathbf{E}\left\{\sum_{t=1}^T \gamma_t^2 \|g_t + \Delta_t\|_*^2\right\} \leq 2 \sum_{t=1}^T \gamma_t^2 [L^2 + \sigma^2]. \quad (10)$$

If (9) is strengthened to

$$\mathbf{E}\{\exp\{\|\Delta_t\|_*^2/\sigma^2\}\} \leq \exp\{1\}, \quad (11)$$

one has

$$\forall \theta > 0 : \text{Prob} \left\{ \sum_{t=1}^T \gamma_t^2 \|g_t + \Delta_t\|_*^2 > 2 \sum_{t=1}^T \gamma_t^2 [L^2 + \sigma^2] + 2\theta\sigma^2 \sum_{t=1}^T \gamma_t^2 \right\} \leq \exp\{1 - \theta\}. \quad (12)$$

**Proof.** (10) is evident due to  $\|g + \Delta\|_*^2 \leq 2[\|g\|_*^2 + \|\Delta\|_*^2]$ . To prove (12), note that the quantity

$$\chi(G(\cdot)) = \inf \{s > 0 : \mathbf{E}\{\exp\{|G|/s\}\} \leq \exp\{1\}\}$$

considered as functional on the space of measurable functions  $G(\cdot)$  on the probability space associated with  $\mathbf{E}\{\cdot\}$  such that  $\chi(G(\cdot))$  is finite, clearly is a norm. It follows that in the case of (11), setting  $a = 2\sigma^2 \sum_{t=1}^T \gamma_t^2$ , we have

$$\mathbf{E}\{\exp\{2 \sum_{t=1}^T \gamma_t^2 \|\Delta_t\|_*^2 / a\}\} \leq \exp\{1\},$$

whence  $\text{Prob}\{[2 \sum_{t=1}^T \gamma_t^2 \|\Delta_t\|_*^2] \geq \theta a\} \leq \exp\{1 - \theta\}$ . Since

$$\sum_{t=1}^T \gamma_t^2 \|g_t + \Delta_t\|_*^2 > 2 \sum_{t=1}^T \gamma_t^2 [L^2 + \sigma^2] + 2\theta \sum_{t=1}^T \gamma_t^2 \sigma^2$$

clearly implies that  $2 \sum_{t=1}^T \gamma_t^2 \|\Delta_t\|_*^2 > 2 \sum_{t=1}^T \gamma_t^2 \sigma^2 + 2\theta\sigma^2 \sum_{t=1}^T \gamma_t^2$  due to  $\|g_t\|_* \leq L$ , (12) follows.  $\square$

**2<sup>0</sup>.** Our next observation is as follows:

**Lemma 1.2** Let  $\Delta_t = H_t(\xi^t) \in E$ , where  $H_t(\cdot)$  are deterministic functions, and  $\xi^t = (\xi_1, \dots, \xi_t)$  with independent  $\xi_t \sim P$ ,  $t = 1, 2, \dots$ , and let  $x_t = X_t(\xi^{t-1}) \in X$ , with deterministic  $X_t(\cdot)$ . Assuming

$$\begin{aligned} (a) \quad & \mathbf{E}_{\xi_t \sim P} \{\|H_t(\xi^t)\|_*^2\} \leq \sigma^2, \quad \forall (t, \xi^{t-1}) \\ (b) \quad & \|\mathbf{E}_{\xi_t \sim P} \{H_t(\xi^t)\}\|_* \leq \mu \quad \forall (t, \xi^{t-1}) \end{aligned} \quad (13)$$

we have for deterministic  $\gamma_t \geq 0$ :

$$\mathbf{E}\left\{\max_{x \in X} \sum_{t=1}^T \gamma_t \langle \Delta_t, x - x_t \rangle\right\} \leq \frac{1}{2}\Omega^2 + \frac{1}{2}\sigma^2 \sum_{t=1}^T \gamma_t^2 + 2\mu\Omega \sum_{t=1}^T \gamma_t. \quad (14)$$

Strengthening (13.a) to

$$\mathbf{E}_{\xi_t \sim P} \left\{ \exp\{\|H_t(\xi^t)\|_*^2 / \sigma^2\} \right\} \leq \exp\{1\}, \quad (15)$$

we get

$$\begin{aligned} \forall(\theta > 0) : \\ \text{Prob} \left\{ \max_{x \in X} \sum_{t=1}^T \gamma_t \langle \Delta_t, x - x_t \rangle \geq \frac{1}{2}\Omega^2 + \frac{1}{2}\sigma^2 \sum_{t=1}^T \gamma_t^2 + 2\mu\Omega \sum_{t=1}^T \gamma_t \right. \\ \left. + \theta \left[ 2\sigma\Omega \sqrt{\sum_{t=1}^T \gamma_t^2} + \sigma^2 \sum_{t=1}^T \gamma_t \right] \right\} \\ \leq \exp\{1 - \theta\} + \exp\{-\theta^2/4\} \end{aligned} \quad (16)$$

**Proof. A.** Let  $y_t = Y_t(\xi^{t-1})$  be given by the recurrence

$$y_1 = y_\omega; y_{t+1} = \text{Prox}_{y_t}(-\gamma_t \Delta_t).$$

Then, same as in the derivation of (3),

$$\forall x \in X : \sum_{t=1}^T \gamma_t \langle -\Delta_t, y_t - x \rangle \leq \frac{1}{2}\Omega^2 + \frac{1}{2} \sum_{t=1}^T \gamma_t^2 \|\Delta_t\|_*^2,$$

so that

$$\forall x \in X : \sum_{t=1}^T \gamma_t \langle \Delta_t, x - x_t \rangle \leq \frac{1}{2}\Omega^2 + \frac{1}{2} \sum_{t=1}^T \gamma_t^2 \|\Delta_t\|_*^2 + \sum_{t=1}^T \gamma_t \langle \Delta_t, y_t - x_t \rangle,$$

whence

$$A := \max_{x \in X} \sum_{t=1}^T \gamma_t \langle \Delta_t, x - x_t \rangle \leq \frac{1}{2}\Omega^2 + \frac{1}{2} \sum_{t=1}^T \gamma_t^2 \|\Delta_t\|_*^2 + \sum_{t=1}^T \gamma_t \langle \Delta_t, y_t - x_t \rangle. \quad (17)$$

Since  $y_t, x_t$  are deterministic functions of  $\xi^{t-1}$  and  $\gamma_t$  are deterministic, we have

$$\mathbf{E}\{\langle \Delta_t, y_t - x_t \rangle\} = \mathbf{E}\{\mathbf{E}_{\xi_t \sim P}\{\langle \Delta_t, y_t - x_t \rangle\}\} = \mathbf{E}\{\langle \mathbf{E}_{\xi_t \sim P}\{\Delta_t\}, y_t - x_t \rangle\} \leq \mathbf{E}\{\mu \|y_t - x_t\|\} \leq 2\mu\Omega$$

(we have used that  $x_t, y_t \in X$ ). Thus, taking expectation of both sides in (17), we get

$$\mathbf{E}\{A\} \leq \frac{1}{2}\Omega^2 + \frac{1}{2}\sigma^2 \sum_{t=1}^T \gamma_t^2 + 2\mu\Omega \sum_{t=1}^T \gamma_t,$$

and (14) follows.

**B.** Now assume that in addition to (13.a) relation (15) takes place, and let us prove (16). Note that (15) implies (13.b) by Jensen's inequality.

**C.** We start with the following observation:

**Lemma 1.3** *Let  $\eta$  be a scalar random variable such that  $|\mathbf{E}\{\eta\}| \leq \nu$  and  $\mathbf{E}\{\exp\{\eta^2\}\} \leq \exp\{1\}$ . Then for all  $\alpha \in \mathbf{R}$  it holds  $\mathbf{E}\{\exp\{\alpha\eta\}\} \leq \exp\{\alpha\nu + \alpha^2\}$ .*

**Proof.** We have  $e^s \leq s + e^{2s^2/3}$  for all  $s$ , whence  $\mathbf{E}\{e^{\alpha\eta}\} \leq \mathbf{E}\{\alpha\eta + e^{2\alpha^2\eta^2/3}\}$ . When  $\alpha^2 \leq 3/2$ , we have  $\mathbf{E}\{e^{2\alpha^2\eta^2/3}\} \leq \exp\{2\alpha^2/3\}$  due to  $\mathbf{E}\{e^{\eta^2}\} \leq \exp\{1\}$  and Moment inequality, so that

$$\mathbf{E}\{e^{\alpha\eta}\} \leq |\alpha| |\mathbf{E}\{\eta\}| + e^{2\alpha^2/3} \leq e^{|\alpha|\nu + 2\alpha^2/3}, \quad 0 \leq \alpha^2 \leq 3/2.$$

Besides this, we have  $\alpha s \leq \frac{1}{4}\alpha^2 + s^2$ , whence  $\mathbf{E}\{e^{\alpha\eta}\} \leq e^{1+\alpha^2/4}$  due to  $\mathbf{E}\{\exp\{\eta^2\}\} \leq e$ . Combining the bounds, we get  $\mathbf{E}\{e^{\alpha\eta}\} \leq e^{|\alpha|\nu + \alpha^2}$  for all  $\alpha$ .  $\square$

**D.** Let  $s_t = \gamma_t \langle \Delta_t, y_t - x_t \rangle$ . Since  $y_t, x_t$  depend solely on  $\xi^{t-1}$  and  $|s_t| \leq 2\gamma_t \|\Delta_t\|_* \Omega$  due to  $x_t, y_t \in X$ , (15) implies that setting  $\sigma_t = 2\gamma_t \sigma \Omega$ , we have

$$\mathbf{E}_{\xi_t \sim P}\{e^{s_t^2/\sigma_t^2}\} \leq \exp\{1\} \quad \forall (t, \xi^{t-1}) \quad (18)$$

Besides this,  $|\mathbf{E}_{\xi_t \sim P}\{s_t\}| = |\gamma_t \langle \mathbf{E}_{\xi_t \sim P}\{\Delta_t\}, y_t - x_t \rangle| \leq 2\gamma_t \Omega \mu$  for all  $t, \xi^{t-1}$ . Applying Lemma 1.3 to the random variable  $\eta = s_t/\sigma_t$ , which allows to set  $\nu = \mu/\sigma$ , we get

$$\mathbf{E}_{\xi_t \sim P}\{e^{\alpha s_t}\} \leq e^{\mu\alpha\sigma_t/\sigma + \alpha^2\sigma_t^2}.$$

Now, for every  $r > 0$  we have

$$\text{Prob}\left\{\sum_{t=1}^T \gamma_t \langle \Delta_t, y_t - x_t \rangle > r\right\} \leq \mathbf{E}\left\{\exp\left\{\alpha \sum_{t=1}^T s_t\right\}\right\} \exp\{-\alpha r\} \forall \alpha \geq 0.$$

Setting  $S_0 = 0$ ,  $S_t = \sum_{\tau=1}^t s_\tau$ , we have for  $\alpha \geq 0$ :

$$\begin{aligned} \mathbf{E}\{\exp\{\alpha S_t\}\} &= \mathbf{E}\{\exp\{\alpha S_{t-1} + \alpha s_t\}\} = \mathbf{E}\{\exp\{\alpha S_{t-1}\} \mathbf{E}_{\xi_t \sim P}\{e^{\alpha s_t}\}\} \\ &\leq \mathbf{E}\{\exp\{\alpha S_{t-1}\} \exp\{\mu \alpha \sigma_t / \sigma + \alpha^2 \sigma_t^2\}\}, \end{aligned}$$

so that

$$\mathbf{E}\{\exp\{\alpha S_T\}\} \leq \exp\left\{\sum_{t=1}^T [\mu \alpha \sigma_t / \sigma + \alpha^2 \sigma_t^2]\right\}$$

and thus

$$\text{Prob}\left\{\sum_{t=1}^T \gamma_t \langle \Delta_t, y_t - x_t \rangle > r\right\} \leq \inf_{\alpha > 0} \left[ \exp\left\{\alpha \left[\sum_{t=1}^T \mu \sigma_t / \sigma - r\right] + \alpha^2 \sum_{t=1}^T \sigma_t^2\right\} \right]. \quad (19)$$

Assuming

$$r = \mu \sum_{t=1}^T \sigma_t / \sigma + \rho = 2\mu\Omega \sum_{t=1}^T \gamma_t + \rho$$

with some positive  $\rho$ , and setting  $\alpha = \frac{\rho}{2\sum_{t=1}^T \sigma_t^2}$ , we get from (19) that

$$\forall \rho > 0 : \text{Prob}\left\{\sum_{t=1}^T \gamma_t \langle \Delta_t, y_t - x_t \rangle > 2\mu\Omega \sum_{t=1}^T \gamma_t + \rho\right\} \leq \exp\left\{-\rho^2 / \left(4 \sum_{t=1}^T \sigma_t^2\right)\right\},$$

or, which is the same,

$$\forall \theta > 0 : \text{Prob}\left\{\sum_{t=1}^T \gamma_t \langle \Delta_t, y_t - x_t \rangle > 2\mu\Omega \sum_{t=1}^T \gamma_t + 2\theta\sigma\Omega \sqrt{\sum_{t=1}^T \gamma_t^2}\right\} \leq \exp\{-\theta^2/4\}. \quad (20)$$

**E.** Acting exactly as in the proof of Lemma 1.1 with  $L$  set to 0, we get

$$\forall \theta > 0 : \text{Prob}\left\{\sum_{t=1}^T \gamma_t^2 \|\Delta_t\|_*^2 > \sigma^2 \sum_{t=1}^T \gamma_t^2 + \theta\sigma^2 \sum_{t=1}^T \gamma_t^2\right\} \leq \exp\{1 - \theta\}. \quad (21)$$



This combines with (20) and (17) to imply (16). Lemma 1.2 is proved.  $\square$

$\mathbf{3}^0$ . Now we can prove Theorem 1.1. Combining (14), (10) and (3), we arrive at (4); (i) is proved. In the case of (5), we have at our disposal both (12) and (16), and these two relations clearly imply item (ii) of Theorem.  $\square$

## 2 Mirror Prox: Stochastic Case

**Situation:** For every  $t = 1, 2, \dots$ , at  $(2t - 1)$ -st call to Stochastic Oracle, the query point being  $x_t \in X$ , the SO returns vector  $g_t(x_t) + \Delta_{2t-1}(x_t, \xi_{2t-1})$ ; at  $(2t)$ -th call, the query point being  $w_t \in X$ , the SO returns  $g_t(w_t) + \Delta_{2t}(w_t, \xi_{2t})$ , with independent  $\xi_s \sim P$ ,  $s = 1, 2, \dots$ . Besides this, we have

$$\begin{aligned} (a) \quad & \|g_t(x) - g_t(x')\|_* \leq \mathcal{M}\|x - x'\| + L \quad \forall (x, x' \in X, t = 1, 2, \dots) \quad [\mathcal{M}, L < \infty], \\ (b) \quad & \mathbf{E}_\xi\{\|\Delta_s(x, \xi)\|_*^2\} \leq \sigma^2, \quad \forall (x \in X, s = 1, 2, \dots) \\ (c) \quad & \|\mathbf{E}_\xi\{\Delta_s(x, \xi)\}\|_* \leq \mu \quad \forall (x \in X, s = 1, 2, \dots) \end{aligned} \tag{22}$$

For the MP recurrence

$$x_1 = x_\omega; w_t = \text{Prox}_{x_t}(\gamma_t[g_t(x_t) + \Delta_{2t-1}(x_t, \xi_{2t-1})]); x_{t+1} = \text{Prox}_{x_t}(\gamma_t[g_t(w_t) + \Delta_{2t}(w_t, \xi_{2t})])$$

we have, setting  $\widehat{g}_t = g_t(x_t)$ ,  $g_t = g_t(w_t)$ ,  $\eta_t = \Delta_{2t-1}(x_t, \xi_{2t-1})$ ,  $\zeta_t = \Delta_{2t}(w_t, \xi_{2t})$ :

$$\begin{aligned} & \gamma_t \langle g_t + \zeta_t, x_{t+1} - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) - V_{x_t}(x_{t+1}) \quad [\text{see (2) in Transparencies}] \\ & \gamma_t \langle \widehat{g}_t + \eta_t, w_t - x_{t+1} \rangle \leq V_{x_t}(x_{t+1}) - V_{w_t}(x_{t+1}) - V_{x_t}(w_t) \quad [\text{see (2) in Transparencies}] \\ \Rightarrow & \gamma_t \langle g_t + \zeta_t, w_t - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) + [\gamma_t \langle g_t + \zeta_t, w_t - x_{t+1} \rangle - V_{x_t}(x_{t+1})] \\ \Rightarrow & \gamma_t \langle g_t + \zeta_t, w_t - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) + \gamma_t \langle g_t + \zeta_t - \widehat{g}_t - \eta_t, w_t - x_{t+1} \rangle + \langle \widehat{g}_t + \eta_t, w_t - x_{t+1} \rangle - V_{x_t}(x_{t+1}) \\ \leq & V_{x_t}(x) - V_{x_{t+1}}(x) + \gamma_t \langle g_t + \zeta_t - \widehat{g}_t - \eta_t, w_t - x_{t+1} \rangle + V_{x_t}(x_{t+1}) - V_{w_t}(x_{t+1}) - V_{x_t}(w_t) - V_{x_t}(x_{t+1}) \\ \Rightarrow & \gamma_t \langle g_t + \zeta_t, w_t - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) + [\gamma_t \langle g_t + \zeta_t - \widehat{g}_t - \eta_t, w_t - x_{t+1} \rangle - V_{w_t}(x_{t+1}) - V_{x_t}(w_t)] \\ \Rightarrow & \gamma_t \langle g_t + \zeta_t, w_t - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) + \gamma_t \|g_t - \widehat{g}_t\|_* \|w_t - x_{t+1}\| + \gamma_t \|\zeta_t - \eta_t\|_* \|w_t - x_{t+1}\| - \frac{1}{2} \|x_t - w_t\|^2 - \frac{1}{2} \|x_{t+1} - w_t\|^2 \end{aligned}$$

Assuming

$$\gamma_t \mathcal{M} \leq 1/2, \quad (23)$$

we have

$$\begin{aligned} & \gamma_t \|g_t - \widehat{g}_t\|_* \|w_t - x_{t+1}\| + \gamma_t \|\zeta_t - \eta_t\|_* \|w_t - x_{t+1}\| - \frac{1}{2} \|x_t - w_t\|^2 - \frac{1}{2} \|x_{t+1} - w_t\|^2 \\ \leq & \gamma_t \mathcal{M} \|w_t - x_t\| \|w_t - x_{t+1}\| + \gamma_t L \|w_t - x_{t+1}\| + \gamma_t \|\zeta_t - \eta_t\|_* \|w_t - x_{t+1}\| \\ & - \frac{1}{2} \|x_t - w_t\|^2 - \frac{1}{2} \|x_{t+1} - w_t\|^2 \\ \leq & \frac{1}{2} [-\|x_t - w_t\|^2 - \frac{1}{4} \|x_{t+1} - w_t\|^2 + \|w_t - x_t\| \|w_t - x_{t+1}\|] \\ & + [\gamma_t [L + \|\zeta_t - \eta_t\|_*] \|w_t - x_{t+1}\| - \frac{3}{8} \|w_t - x_{t+1}\|^2] \\ \leq & \gamma_t^2 [L + \|\zeta_t\|_* + \|\eta_t\|_*]^2 \leq 3\gamma_t^2 [L^2 + \|\zeta_t\|_*^2 + \|\eta_t\|_*^2] \end{aligned}$$

Thus, for every  $x \in X$  and every  $t$  we have

$$\gamma_t \langle g_t + \zeta_t, w_t - x \rangle \leq V_{x_t}(x) - V_{x_{t+1}}(x) + 3\gamma_t^2 [L^2 + \|\zeta_t\|_*^2 + \|\eta_t\|_*^2],$$

whence

$$\sum_{t=1}^T \gamma_t \langle g_t, w_t - x \rangle \leq \frac{1}{2} \Omega^2 + 3 \sum_{t=1}^T \gamma_t^2 [L^2 + \|\zeta_t\|_*^2 + \|\eta_t\|_*^2] + \sum_{t=1}^T \gamma_t \langle \zeta_t, x - w_t \rangle.$$

Therefore with  $\lambda_t := \lambda_t^T = \gamma_t / \sum_{s=1}^T \gamma_s$  we have

$$\epsilon_T := \max_{x \in X} \sum_{t=1}^T \lambda_t \langle g_t, w_t - x \rangle \leq \frac{\frac{1}{2} \Omega^2 + 3 \sum_{t=1}^T \gamma_t^2 [L^2 + \|\zeta_t\|_*^2 + \|\eta_t\|_*^2] + \max_{x \in X} \sum_{t=1}^T \gamma_t \langle \zeta_t, x - w_t \rangle}{\sum_{\tau=1}^T \gamma_\tau} \quad (24)$$

**Our goal** is to prove the following

**Theorem 2.1** (i) *Assuming (22), (23), one has for every  $T = 1, 2, \dots$*

$$\mathbf{E} \{ \epsilon_T \} \leq \Xi := \frac{\Omega^2 + [3L^2 + 13\sigma^2/2] \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t} + 2\mu\Omega. \quad (25)$$

(ii) *Strengthening (22.b) to*

$$\mathbf{E}_\xi \{ \exp\{ \|\Delta_s(x, \xi)\|_*^2 / \sigma^2 \} \} \leq \exp\{1\} \quad \forall (x \in X, s = 1, 2, \dots) \quad (26)$$

we have for every  $\theta > 0$ :

$$\begin{aligned} \text{Prob}\{\epsilon_T > \Xi + \theta\Upsilon\} &\leq 6 \exp\{-\theta\} + \exp\{-\theta^2/4\}, \\ \Upsilon &= \frac{7\sigma^2 \sum_{t=1}^T \gamma_t^2 + 2\sigma\Omega\sqrt{\sum_{t=1}^T \gamma_t^2}}{\sum_{t=1}^T \gamma_t}. \end{aligned} \quad (27)$$

Note that with the stepsizes

$$\gamma_t = \min \left[ \frac{1}{2\mathcal{M}}, \frac{\Omega}{\sqrt{3L^2 + 13\sigma^2/2\sqrt{T}}} \right], \quad 1 \leq t \leq T \quad (28)$$

one has

$$\Xi = \frac{2\Omega^2\mathcal{M}}{T} + \frac{2\Omega\sqrt{3L^2 + 13\sigma^2/2}}{\sqrt{T}} + 2\mu\Omega, \quad \Upsilon \leq 9\frac{\Omega\sigma}{\sqrt{T}}. \quad (29)$$

**Proof** repeats word by word the one of Theorem 1.1, with (24) in the role of (3).

### 3 Proximal Setup for $\ell_1/\ell_2$ Ball

Let

$$X = \{x = [x^1; \dots; x^n] \subset E = \mathbf{R}^{k_1} \times \dots \times \mathbf{R}^{k_n} : \sum_{j=1}^n \|x^j\|_2 \leq 1\}$$

and

$$\omega(x) = \frac{1}{p\gamma} \sum_{j=1}^p \|x^j\|_2^p, \quad p = \begin{cases} 2, & n \leq 2 \\ 1 + \frac{1}{\ln n}, & n \geq 3 \end{cases}, \quad \gamma = \begin{cases} 1, & n = 1 \\ \frac{1}{2}, & n = 2 \\ \frac{1}{e \ln(n)}, & n > 2 \end{cases}$$

We have for  $x \in X' = \{x \in X : x^j \neq 0 \forall j\}$ :

$$\begin{aligned}
\gamma D\omega(x)[h] &= \sum_{j=1}^n \|x^j\|_2^{p-2} \langle x^j, h^j \rangle \\
\gamma D^2\omega(x)[h, h] &= -(2-p) \sum_{j=1}^n \|x^j\|_2^{p-4} [\langle x^j, h^j \rangle]^2 + \sum_{j=1}^n \|x^j\|_2^{p-2} \|h^j\|_2^2 \\
&\geq \sum_{j=1}^n \|x^j\|_2^{p-2} \|h^j\|_2^2 - (2-p) \sum_{j=1}^n \|x^j\|_2^{p-4} \|x^j\|_2^2 \|h^j\|_2^2 \\
&\geq (p-1) \sum_{j=1}^n \|x^j\|_2^{p-2} \|h^j\|_2^2 \\
&\Rightarrow \left[ \sum_j \|h^j\|_2 \right]^2 = \left[ \sum_{j=1}^n [\|h^j\|_2 \|x^j\|_2^{\frac{p-2}{2}}] \|x^j\|_2^{\frac{2-p}{2}} \right]^2 \leq \left[ \sum_{j=1}^n \|h^j\|_2^2 \|x^j\|_2^{p-2} \right] \left[ \sum_{j=1}^n \|x^j\|_2^{2-p} \right] \\
&\Rightarrow \left[ \sum_j \|h^j\|_2 \right]^2 \leq \left[ \sum_{j=1}^n \|x^j\|_2^{2-p} \right] \frac{\gamma}{p-1} D^2\omega(x)[h, h]
\end{aligned}$$

Setting  $t_j = \|x^j\|_2 \geq 0$ , we have  $\sum_j t_j \leq 1$ , whence due to  $0 \leq 2-p \leq 1$  it holds  $\sum_j t_j^{2-p} \leq n n^{-(2-p)} = n^{p-1}$ . Thus,

$$\left[ \sum_j \|h^j\|_2 \right]^2 \leq n^{p-1} \frac{\gamma}{p-1} D^2\omega(x)[h, h]$$

while

$$\max_{x \in X} \omega(x) - \min_{x \in X} \omega(x) \leq \frac{1}{\gamma p} \tag{30}$$

With  $p, \gamma$  as above, when  $n \geq 3$  we get  $\frac{\gamma}{p-1} n^{p-1} = \frac{1}{e \ln(n)/\ln(n)} n^{1/\ln(n)} = 1$ , and similarly for  $n = 1, 2$ . Consequently,

$$\forall (x \in X', h) : \left[ \sum_{j=1}^n \|h^s_j\|_2 \right]^2 \leq D^2\omega(x)[h, h]. \tag{31}$$

Since  $\omega(\cdot)$  is continuously differentiable and the complement of  $X'$  in  $X$  is the union of finitely many proper linear subspaces of  $E$ , (31) implies that  $\omega$  is strongly convex on  $X$ , modulus 1, w.r.t. the  $\ell_1/\ell_2$  norm. Besides this, we have

$$\frac{1}{\gamma p} = \left\{ \begin{array}{ll} \frac{1}{2}, & n = 1 \\ 2, & n = 2 \\ \leq e \ln(n), & n \geq 3 \end{array} \right\} \leq O(1) \ln(n+1).$$

which combines with (30) to imply that the  $\omega$ -radius of  $X$  is  $\leq O(1) \sqrt{\ln(n+1)}$ .

## 4 Proximal Setup for Nuclear Norm Ball

For  $y \in \mathbf{S}^n$ , let  $\lambda(y)$  be the vector of eigenvalues of  $y$  (taken with their multiplicities in the non-ascending order), and let  $|y|_1 = \|\lambda(y)\|_1$  be the trace norm.

**Proposition 4.1** *Let  $N \geq M \geq 3$ , and let  $E$  be a linear subspace in  $\mathbf{S}^N$  such that every matrix  $y \in E$  has at most  $M$  nonzero eigenvalues. Let  $q = \frac{1}{\ln(M)}$ , so that  $0 < q < 1$ , and let*

$$\widehat{\omega}(y) = \frac{4e \ln(M)}{1+q} \sum_{j=1}^N |\lambda_j(y)|^{1+q} : \mathbf{S}^N \rightarrow \mathbf{R}.$$

*The function  $\widehat{\omega}(\cdot)$  is continuously differentiable, convex, and its restriction on the set  $Y_E = \{y \in E : |y|_1 \leq 1\}$  is strongly convex, modulus 1, w.r.t.  $|\cdot|_1$ . Besides this,*

$$\forall (y \in Y, h \in \mathbf{S}^N) : |\langle \omega'(y), h \rangle| \leq 4e \ln(M) |h|_1. \quad (32)$$

**Proof. 1<sup>0</sup>.** Let  $0 < q < 1$ . Consider the following function of  $y \in \mathbf{S}^N$ :

$$\chi(y) = \frac{1}{1+q} \sum_{i=1}^N |\lambda_i(y)|^{1+q} = \text{Tr}(f(y)), \quad f(s) = \frac{1}{1+q} |s|^{1+q}.$$

**2<sup>0</sup>.** Function  $f(s)$  is continuously differentiable on the axis and twice continuously differentiable outside of the origin; consequently, we can find a sequence of polynomials  $f_k(s)$  converging, as  $k \rightarrow \infty$ , to  $f$  along with their first derivatives uniformly on every compact subset of  $\mathbf{R}$  and, besides this, converging to  $f$  uniformly along with the first and the second derivative on every compact subset of  $\mathbf{R} \setminus \{0\}$ . Now let  $y, h \in \mathbf{S}^N$ , let  $y = u \text{Diag}\{\lambda\} u^T$  be the eigenvalue decomposition of  $y$ , and let  $h = \widehat{u} \widehat{h} u^T$ . For a polynomial  $p(s) = \sum_{k=0}^K p_k s^k$ , setting

$P(w) = \text{Tr}(\sum_{k=0}^K p_k w^k) : \mathbf{S}^N \rightarrow \mathbf{R}$ , and denoting by  $\gamma$  a closed contour in  $\mathbb{C}$  encircling the spectrum of  $y$ , we have

$$\begin{aligned}
(a) \quad & P(y) = \text{Tr}(p(y)) = \sum_{j=1}^N p(\lambda_j(y)) \\
(b) \quad & DP(y)[h] = \text{Tr}(\sum_{k=0}^K k p_k \text{Tr}(y^{k-1}h)) = \text{Tr}(p'(y)h) = \sum_{j=1}^N p'(\lambda_j(y)) \widehat{h}_{jj} \\
(c) \quad & D^2P(y)[h, h] = \left. \frac{d}{dt} \right|_{t=0} DP(y+th)[h] = \left. \frac{d}{dt} \right|_{t=0} \text{Tr}(p'(y+th)h) \\
& = \left. \frac{d}{dt} \right|_{t=0} \frac{1}{2\pi i} \oint_{\gamma} \text{Tr}(h(zI - (y+th))^{-1}) p'(z) dz = \frac{1}{2\pi i} \oint_{\gamma} \text{Tr}(h(zI - y)^{-1} h(zI - y)^{-1}) p'(z) dz \\
& = \frac{1}{2\pi i} \oint_{\gamma} \sum_{i,j=1}^N \widehat{h}_{ij}^2 \frac{p'(z)}{(z-\lambda_i(y))(z-\lambda_j(y))} dz = \sum_{i,j=1}^N \widehat{h}_{ij}^2 \Gamma_{ij}, \\
& \Gamma_{ij} = \begin{cases} \frac{p'(\lambda_i(y)) - p'(\lambda_j(y))}{\lambda_i(y) - \lambda_j(y)}, & \lambda_i(y) \neq \lambda_j(y) \\ p''(\lambda_i(y)), & \lambda_i(y) = \lambda_j(y) \end{cases}
\end{aligned}$$

We conclude from (a, b) that as  $k \rightarrow \infty$ , the real-valued polynomials  $F_k(\cdot) = \text{Tr}(f_k(\cdot))$  on  $\mathbf{S}^N$  converge, along with their first order derivatives, uniformly on every bounded subset of  $\mathbf{S}^N$ , and the limit of the sequence, by (a), is exactly  $\chi(\cdot)$ . Thus,  $\chi(\cdot)$  is continuously differentiable, and (b) says that

$$D\chi(y)[h] = \sum_{j=1}^N f'(\lambda_j(y)) \widehat{h}_{jj}. \quad (33)$$

Besides this, (a-c) say that if  $U$  is a closed convex set in  $\mathbf{S}^N$  which does not contain singular matrices, then  $F_k(\cdot)$ , as  $k \rightarrow \infty$ , converge along with the first and the second derivative uniformly on every compact subset of  $U$ , so that  $\chi(\cdot)$  is twice continuously differentiable on  $U$ , and at every point  $y \in U$  we have

$$D^2\chi(y)[h, h] = \sum_{i,j=1}^N \widehat{h}_{ij}^2 \Gamma_{ij}, \quad \Gamma_{ij} = \begin{cases} \frac{f'(\lambda_i(y)) - f'(\lambda_j(y))}{\lambda_i(y) - \lambda_j(y)}, & \lambda_i(y) \neq \lambda_j(y) \\ f''(\lambda_i(y)), & \lambda_i(y) = \lambda_j(y) \end{cases} \quad (34)$$

and in particular  $\chi(\cdot)$  is convex on  $U$ .

**3<sup>0</sup>.** We intend to prove that (i)  $\chi(\cdot)$  is convex, and (ii) its restriction on the unit ball  $Y$  of the trace norm is strongly convex, with certain modulus  $\alpha > 0$ , w.r.t. the trace norm  $|\cdot|_1$ . Since  $\chi$  is continuously differentiable, all we need to prove (i) is to verify that

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \geq 0 \quad (*)$$

for a dense in  $\mathbf{S}^n \times \mathbf{S}^N$  set of pairs  $(y', y'')$ , e.g., those with nonsingular  $y' - y''$ . For a pair of the latter type, the polynomial  $q(t) = \text{Det}(y' + t(y'' - y'))$  of  $t \in \mathbf{R}$  is not identically zero and thus has finitely many roots on  $[0, 1]$ . In other words, we can find finitely many points  $t_0 = 0 < t_1 < \dots < t_n = 1$  such that all “matrix intervals”  $\Delta_i = (y_i, y_{i+1})$ ,  $y_k = y' + t_k(y'' - y')$ ,  $1 \leq i \leq n - 1$ , are comprised of nonsingular matrices. Therefore  $\chi$  is convex on every compact subset of every interval  $\Delta_i$ , and since  $\chi$  is continuously differentiable, (\*) follows.

4<sup>0</sup>. Now let us prove that with properly defined  $\alpha > 0$  one has

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \geq \alpha |y' - y''|_1^2 \quad \forall y', y'' \in Y_E$$

Let  $\epsilon > 0$ , and let  $Y^\epsilon$  be a convex open in  $Y = \{y : |y|_1 \leq 1\}$  neighbourhood of  $Y_E$  such that for all  $y \in Y^\epsilon$  at most  $M$  eigenvalues of  $y$  are of magnitude  $> \epsilon$ . We intend to prove that for some  $\alpha_\epsilon > 0$  one has

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \geq \alpha_\epsilon |y' - y''|_1^2 \quad \forall y', y'' \in Y^\epsilon. \quad (35)$$

Same as above, it suffices to verify this relation for a dense in  $Y^\epsilon \times Y^\epsilon$  set of pairs  $y', y'' \in Y^\epsilon$ , e.g., for those pairs  $y', y'' \in Y^\epsilon$  for which  $y' - y''$  is nonsingular. Defining matrix intervals  $\Delta_i$  as above and taking into account continuous differentiability of  $\chi$ , it suffices to verify that if  $y \in \Delta_i$  and  $h = y' - y''$ , then  $D^2\chi(y)[h, h] \geq \alpha_\epsilon |h|_1^2$ . To this end observe that by (34) all we have to prove is that

$$D^2\chi(y)[h, h] = \sum_{i,j=1}^N \hat{h}_{ij}^2 \Gamma_{ij} \geq \alpha_\epsilon |h|_1^2. \quad (\#)$$

Setting  $\lambda_j = \lambda_j(y)$ , observe that  $\lambda_i \neq 0$  for all  $i$  due to the origin of  $y$ , and if  $|\lambda_i| \geq |\lambda_j|$ , then  $\Gamma_{ij} \geq q|\lambda_i|^{q-1}$ . Indeed, the latter relation definitely holds true when  $\lambda_i = \lambda_j$ . Now, if  $\lambda_i$  and  $\lambda_j$  are of the same sign, then  $\Gamma_{ij} = \frac{|\lambda_i|^{q-|\lambda_j|^q}}{|\lambda_i| - |\lambda_j|} \geq q|\lambda_i|^{q-1}$ , since the derivative of the concave (recall that  $0 < q \leq 1$ ) function  $t^q$  of  $t > 0$  is positive and nonincreasing. If  $\lambda_i$  and  $\lambda_j$  are of different signs, then  $\Gamma_{ij} = \frac{|\lambda_i|^{q+|\lambda_j|^q}}{|\lambda_i| + |\lambda_j|} \geq |\lambda_i|^{q-1}$  due to  $|\lambda_j|^q \geq |\lambda_j||\lambda_i|^{q-1}$ , and therefore  $\Gamma_{ij} \geq q|\lambda_i|^{q-1}$ . Without loss of generality, we can assume that the positive reals  $\mu_i = |\lambda_i|$ ,  $i = 1, \dots, N$ , form a nondecreasing sequence, so that, by above,  $\Gamma_{ij} \geq q\mu_j^{q-1}$  when  $i \leq j$ . Besides this, at most  $M$  of  $\mu_j$  are  $\geq \epsilon$ , since  $y', y'' \in Y^\epsilon$  and therefore  $y \in Y^\epsilon$  by convexity of  $Y^\epsilon$ . By the above,

$$D^2\chi(y)[h, h] = 2q \sum_{i < j \leq N} \hat{h}_{ij}^2 \mu_j^{q-1} + q \sum_{j=1}^N \hat{h}_{jj}^2 \mu_j^{q-1},$$

or, equivalently by symmetry of  $\widehat{h}$ , if

$$h^j = \begin{bmatrix} & & & \widehat{h}_{1j} \\ & & & \widehat{h}_{2j} \\ & & & \vdots \\ \widehat{h}_{j1} & \widehat{h}_{j2} & \cdots & \widehat{h}_{jj} \end{bmatrix}$$

and  $H_j$  is the Frobenius norm of  $h^j$ , then

$$D^2\chi(y)[h, h] = q \sum_{j=1}^N H_j^2 \mu_j^{q-1} \geq q\epsilon^{q-1} \sum_{j=1}^{N-M} H_j^2 + q \sum_{j=N-M+1}^N H_j^2 \mu_j^{q-1}.$$

Now note that  $\mu_j > 0$  and  $\sum_{j=N-M+1}^N \mu_j \leq 1$  due to  $y \in Y$ . It follows that setting  $\eta = [H_{N-M+1}; H_{N-M+2}; \dots; H_N]$ , we have

$$\begin{aligned} \sum_{j=N-M+1}^N H_j^2 \mu_j^{q-1} &\geq \min_{\nu_j > 0: \sum_{j=N-M+1}^N \nu_j \leq 1} \sum_{j=N-M+1}^N H_j^2 \nu_j^{q-1} = \left[ \sum_{j=N-M+1}^N H_j^{\frac{2}{2-q}} \right]^{2-q} \\ &= \|\eta\|_{\frac{2}{2-q}}^2 \geq M^{-2[1-\frac{2-q}{2}]} \|\eta\|_1^2 = M^{-q} \|\eta\|_1^2, \end{aligned}$$

(when computing the minimum, take into account that  $0 < q < 1$ ). Besides this, setting  $\zeta = [H_1; H_2; \dots; H_{N-M}]$ , we have

$$\|\zeta\|_1^2 \leq (N-M) \sum_{j=1}^{N-M} H_j^2 \leq [\epsilon^{1-q}(N-M)]\epsilon^{q-1} \sum_{j=1}^{N-M} H_j^2$$

We see that for every positive  $\delta$  one has

$$\begin{aligned} \left[ \sum_{j=1}^N H_j \right]^2 &= [\|\zeta\|_1 + \|\eta\|_1]^2 \leq (1+\delta)\|\eta\|_1^2 + (1+\delta^{-1})\|\zeta\|_1^2 \\ &\leq (1+\delta)M^q \sum_{j=N-M+1}^N H_j^2 \mu_j^{q-1} + (1+\delta^{-1})[\epsilon^{1-q}(N-M)]\epsilon^{q-1} \sum_{j=1}^{N-M} H_j^2 \\ &\leq \max[(1+\delta)M^q, (1+\delta^{-1})\epsilon^{1-q}(N-M)] \left[ \epsilon^{1-q} \sum_{j=1}^{N-M} H_j^2 + \sum_{j=N-M+1}^N H_j^2 \mu_j^{q-1} \right] \\ &\leq q^{-1} \max[(1+\delta)M^q, (1+\delta^{-1})\epsilon^{1-q}(N-M)] D^2\chi(y)[h, h]. \end{aligned}$$



Now observe that  $\widehat{h} = \sum_{j=1}^N h^j$  and  $h^j$  is of rank  $\leq 2$ , so that  $|h^j|_1$  is at most twice the Frobenius norm  $H_j$  of  $h^j$ . Therefore

$$|h|_1^2 = |\widehat{h}|_1^2 \leq 4 \left[ \sum_{j=1}^N H_j \right]^2 \leq 4q^{-1} \max[(1 + \delta)M^q, (1 + \delta^{-1})\epsilon^{1-q}(N - M)] D^2 \chi(y)[h, h].$$

This inequality holds true for all  $\delta$ . Setting

$$\alpha_\epsilon^{-1} = \min_{\delta > 0} 4q^{-1} \max[(1 + \delta)M^q, (1 + \delta^{-1})\epsilon^{1-q}(N - M)],$$

we ensure the validity of (#), and consequently the validity of (35). The latter relation, combined with  $\alpha_\epsilon \rightarrow \alpha = qM^{-q}/4$  as  $\epsilon \rightarrow +0$  due to  $q < 1$ , implies that

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \geq \alpha |y' - y''|_1^2 \quad \forall (y', y'' \in Y_E), \quad \alpha = qM^{-q}/4.$$

Setting  $q = \frac{1}{\ln(M)}$  and observing that with this  $q$ ,  $\alpha = [4e \ln(M)]^{-1}$ , so that  $\widehat{\omega}(\cdot) = \alpha^{-1} \chi(\cdot)$ , we see that  $\widehat{\omega}$  indeed is continuously differentiable convex function on  $\mathbf{S}^N$  which is strongly convex, modulus 1 w.r.t.  $|\cdot|_1$ , on  $Y_E$ . It remains to note that by (33) for  $y \in Y$  and  $h \in \mathbf{S}^N$  we have

$$\begin{aligned} |\langle \omega'(y), h \rangle| &= 4e \ln(M) |\langle \chi'(y), h \rangle| \leq 4e \ln(M) \sum_{j=1}^N |\lambda_j(y)|^q |\widehat{h}_{jj}| \\ &\leq 4e \ln(M) \sum_j |\widehat{h}_{jj}| \leq 4e \ln(M) |\widehat{h}|_1 = 4e \ln(M) |h|_1. \end{aligned} \quad \square$$

Now let  $m, n$  be positive integers with  $2 \leq m \leq n$ , and let  $N = m + n$ ,  $M = 2m$ . For  $x \in \mathbf{R}^{m \times n}$ , let  $\sigma_i(x)$ ,  $1 \leq i \leq m$ , be the singular values of  $x$ , let  $\|x\|_{\text{nuc}}$  be the nuclear norm of  $x$ , and let  $\mathcal{A}x = \frac{1}{2} \begin{bmatrix} x & x^T \end{bmatrix} \in \mathbf{S}^N$ . Observe that the image space  $E$  of  $\mathcal{A}$  is a linear subspace of  $\mathbf{S}^N$ , and that the eigenvalues of  $y = \mathcal{A}x$  are the  $2m$  reals  $\pm \sigma_i(x)/2$ ,  $1 \leq i \leq m$ , and  $N - m$  zeros, so that  $\|x\|_{\text{nuc}} \equiv |\mathcal{A}x|_1$  and  $M = 2m, E$  satisfy the premise of Proposition 4.1. Setting

$$\omega(x) = \widehat{\omega}(\mathcal{A}x) = \frac{4e \ln(2m)}{2^q(1 + q)} \sum_i \sigma_i^{1+q}(x), \quad q = \frac{1}{\ln(2m)},$$

and invoking Proposition 4.1, we see that  $\omega$  is a convex continuously differentiable function on  $\mathbf{R}^{m \times n}$  which, due to the identity  $\|x\|_{\text{nuc}} \equiv |\mathcal{A}x|_1$ , is strongly convex, modulus 1 w.r.t.  $\|\cdot\|_{\text{nuc}}$ , on the  $\|\cdot\|_{\text{nuc}}$ -unit ball  $X$ . Observe that

$$\Omega[X, \omega(\cdot)] \leq 2\sqrt{2e \ln(2m)} \leq 5\sqrt{\ln(2m)}.$$

## 5 Mirror Descent in Semi-Bounded Case

**Theorem 5.1** *Let  $\|\cdot\|, \omega$  be a proximal setup for  $X \subset E$ , and assume that  $X$  contains  $\|\cdot\|$ -ball of positive radius  $\theta\Omega$  centered at some point  $c$ . Consider MD trajectory*

$$x_1 = x_\omega; x_{\tau+1} = \text{Prox}_{x_\tau}(\gamma_\tau g_\tau(x_\tau)) \quad (36)$$

with stepsizes

$$\gamma_\tau = \begin{cases} \nu_\tau / \|g_\tau(x_\tau)\|_*, & g_\tau(x_\tau) \neq 0 \\ \nu_\tau, & g_\tau(x_\tau) = 0 \end{cases} \quad (37)$$

Assume the vector fields  $g_\tau(\cdot)$  are uniformly semi-bounded on  $X$ :

$$\sup_{x, x' \in X, \tau} \langle g_\tau(x), x' - x \rangle \leq M < \infty \quad (38)$$

Then with  $x^t$  defined as

$$x^t = \sum_{\tau=1}^t \lambda_\tau^t x_\tau, \quad \lambda_\tau^t = \gamma_\tau / \sum_{s=1}^t \gamma_s$$

when  $g_\tau(x_\tau) \neq 0$  for all  $\tau \leq t$ , otherwise defined as (any)  $x_\tau$  such that  $g_\tau(x_\tau) = 0$ , the following holds true: when

$$\Delta_t := \frac{\Omega^2 + \sum_{\tau=1}^t \nu_\tau^2}{2 \sum_{\tau=1}^t \nu_\tau} < \theta\Omega,$$

one has

$$\epsilon_t := \max_{x \in X} \sum_{\tau=1}^t \lambda_\tau^t \langle g_\tau(x_\tau), x_\tau - x \rangle \leq \frac{M \Delta_t}{\theta\Omega - \Delta_t}. \quad (39)$$

In particular, when  $t \geq 4\theta^{-2}$  and  $\nu_\tau = \frac{\Omega}{\sqrt{t}}$ ,  $1 \leq \tau \leq t$ , one has

$$\epsilon_t \leq \frac{2M}{\theta\sqrt{t}}.$$

**Proof.** There is nothing to prove when  $x^t = x_\tau$  such that  $g_\tau(x_\tau) = 0$ ; thus assume that  $g_\tau(x_\tau) \neq 0$  for all  $\tau \leq t$ . Let  $h_\tau(x) = g_\tau(x) / \|g_\tau(x)\|_*$  when  $g_\tau(x) \neq 0$ , and  $h_\tau(x) = 0$  when  $g_\tau(x) = 0$ . Then the recurrence (36) reads

$$x_1 = x_\omega; x_{\tau+1} = \text{Prox}_{x_\tau}(\nu_\tau h_\tau(x_\tau)) \quad (40)$$

and  $\|h_\tau(x_\tau)\|_* \leq 1$ , whence

$$\max_{x \in X} \sum_{\tau=1}^t \mu_\tau \langle h_\tau(x_\tau), x_\tau - x \rangle \leq \Delta_t := \frac{\Omega^2 + \sum_{\tau=1}^t \nu_\tau^2}{2 \sum_{\tau=1}^t \nu_\tau}, \quad (41)$$

$$\mu_\tau = \nu_\tau / \sum_{s=1}^t \nu_s$$

or

$$\max_{x \in X} \sum_{\tau \leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} \langle g_\tau(x_\tau), x_\tau - x \rangle \leq \Delta_t. \quad (42)$$

Assuming w.l.o.g.  $c = 0$ , we have by (38)

$$\forall (x, \|x\| \leq r := \theta\Omega, \tau) : \langle g_\tau(x_\tau), x - x_\tau \rangle \leq M,$$

whence

$$\|g_\tau(x_\tau)\|_* \leq r^{-1} [M + \langle g_\tau(x_\tau), x_\tau \rangle] \forall \tau$$

or, equivalently,

$$\forall (\tau \leq t) : \frac{1}{\|g_\tau(x_\tau)\|_*} \geq \frac{r}{M} - \frac{1}{M} \frac{\langle g_\tau(x_\tau), x_\tau \rangle}{\|g_\tau(x_\tau)\|_*}$$

and therefore

$$\sum_{\tau \leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} \geq \frac{r}{M} \sum_{\tau \leq t} \mu_\tau - \frac{1}{M} \sum_{\tau \leq t} \mu_\tau \frac{\langle g_\tau(x_\tau), x_\tau \rangle}{\|g_\tau(x_\tau)\|_*}.$$

In other words,

$$\frac{r}{M} \leq \sum_{\tau \leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} + \frac{1}{M} \sum_{\tau \leq t} \mu_\tau \frac{\langle g_\tau, x_\tau \rangle}{\|g_\tau(x_\tau)\|_*} \leq \sum_{\tau \leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} + \frac{1}{M} \Delta_t,$$

where the concluding  $\leq$  is due to (42) and  $0 \in X$ . We see that

$$\sum_{\tau \leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} \geq \frac{r - \Delta_t}{M}.$$

Assuming the right hand side in this inequality positive and taking into account that

$$\lambda_\tau^t = \frac{\mu_\tau / \|g_\tau(x_\tau)\|_*}{\sum_{s \leq t} \mu_s / \|g_s(x_s)\|_*},$$

we get from (42)

$$\max_{x \in X} \sum_{\tau \leq t} \lambda_{\tau}^t \langle g_{\tau}(x_{\tau}), x_{\tau} - x \rangle \leq \frac{M \Delta_t}{r - \Delta_t},$$

as claimed in (39). □