Semi-Proximal Mirror Prox for Nonsmooth Composite Minimization

Niao He
ISyE, Georgia Tech

Joint work with Zaid Harchaoui (NYU & Inria)

ISMP, Pittsburgh, July 17th, 2015

Outline

First-order methods for composite minimization

$$\min_{x\in X} f(x) + h(x)$$

f and h are convex, f **is smooth,** h **is simple**.

- **(Acc)Proximal gradient methods** (when h *proximal-friendly*)
  - Proximal operator:

$$\mathrm{prox}_h(\eta) = \mathrm{argmin}_{x\in X}\{\tfrac{1}{2}\|x-\eta\|_2^2 + h(x)\}$$

  - For example, when $h(x) = \|x\|_1$, reduces to soft thresholding.
  - Worst complexity bound for first-order oracles is $\mathcal{O}(1/\sqrt{\epsilon})$.

- **Conditional gradient methods** (when h is *LMO-friendly*)
  - (Composite) linear minimization oracles(LMO):

$$\mathrm{LMO}_h(\eta) = \mathrm{argmin}_{x\in X}\{\langle \eta, x\rangle + h(x)\}$$

  - For example, when $h(x) = \|x\|_{\mathrm{nuc}}$ or $\delta_{\|x\|_{\mathrm{nuc}}\le 1}(x)$, reduces to computing top pair of singular vectors.
  - Worst (also optimal) complexity bound for LMOs is $\mathcal{O}(1/\epsilon)$.

Structured nonsmooth composite minimization

$$\min_{x \in X} f(x) + h(x)$$

f enjoys the *Fenchel-type representation*

$$f(x) = \max_{z \in Z} \{\langle x, Az \rangle - g(z)\}$$

for some convex function g and convex compact set Z.

Existing LMO-based algorithms are limited to

- **Nesterov's smoothing technique when (Z, g) are simple**
  - Run conditional gradient with the smooth approximation f^μ, where $|f^\mu - f| \leq O(\mu)$. [Lan'13] [PHM'14] [LZ'14]
- **Dual approach when Z admits favorable geometry**
  - Run first-order method on the dual problem [CJN'13] [JN'14]

Worst (also optimal) complexity bound for LMOs is $\mathcal{O}(1/\epsilon^2)$.

Problem of interest

$$\min_{x \in X} \quad f(x) + h_1(x) + h_2(x)$$

- f (perhaps nonsmooth) admits Fenchel-type representation

$$f(x) = \max_{z \in Z} \Phi(x, z)$$

(**without** assuming Z to be simple or have favorable geometry)

- h_1 (perhaps nonsmooth) is proximal-friendly

$$\mathrm{prox}_{h_1}(\eta) = \mathrm{argmin}_{x \in X}\{\tfrac{1}{2}\|x - \eta\|_2^2 + h_1(x)\}$$

- h_2 (perhaps nonsmooth) is LMO-friendly

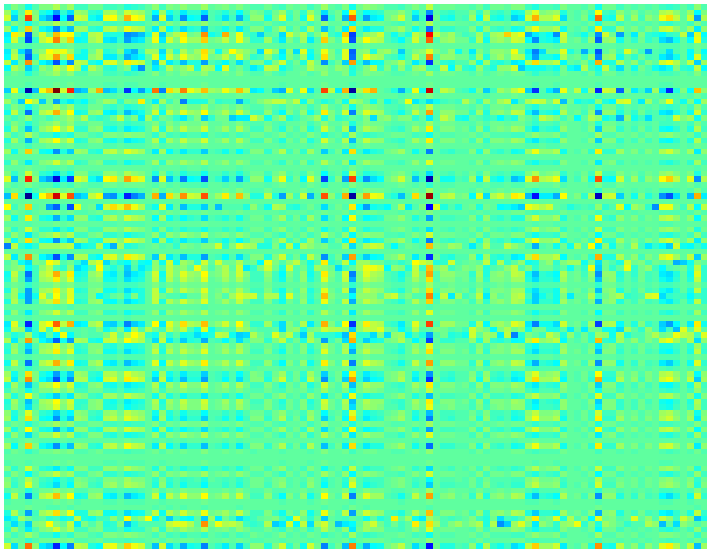
$$\mathrm{LMO}_{h_2}(\eta) = \mathrm{argmin}_{x \in X}\{\langle \eta, x \rangle + h_2(x)\}$$

Example: social network link prediction

$$\min_{x\in\mathbf{R}^{n\times n}} \underbrace{\frac{1}{|E|}\sum_{(i,j)\in E}\max(1-b_{ij}x_{ij},0)}_{f(x)} + \underbrace{\mu_1\|x\|_1}_{h_1(x)} + \underbrace{\mu_2\|x\|_{\text{nuc}}}_{h_2(x)}$$

- $f(x)$ = empirical hinge loss
- $h_1(x) = \ell_1$ norm $\rightarrow$ sparsity
- $h_2(x)$ = nuclear norm $\rightarrow$ low-rank

Social Network



Low-rank and Sparse Matrix

The goal and the cure

To design an algoirthm that solves the above problem

$$\min_{x \in X} \quad f(x) + h_1(x) + h_2(x)$$

by taking advantages and leveraging

- saddle point representation of f
- proximal operators of h_1
- linear minimization oracles of h_2

We propose Semi-Proximal Mirror-Prox that blends *proximal gradient* and *conditional gradient* and achieves best of both worlds

- optimal $\mathcal{O}(1/\epsilon)$ complexity bound in first-order oracles
- optimal $\mathcal{O}(1/\epsilon^2)$ complexity bound in linear minimization oracles

Course of action: saddle point reformulation

$$\min_{x\in X} f(x) + h_1(x) + h_2(x)$$

$$\Updownarrow$$

$$\min_{x,y\in X}\max_{z\in Z} \{\Phi(x,z) + h_1(x) + h_2(y) : x = y\}$$

$$\Updownarrow \text{ (with proper } \rho > 0)$$

$$\min_{x,y\in X}\max_{z\in Z} \Phi(x,z) + h_1(x) + h_2(y) + \rho\|x-y\|_*$$

$$\Updownarrow$$

$$\min_{\substack{x,y\in X\\ \tau_1,\tau_2}}\max_{\substack{z\in Z\\ \|w\|\le 1}} \Phi(x,z) + \rho\langle x-y, w\rangle + \tau_1 + \tau_2$$
$$\tau_1 \ge h_1(x), \tau_2 \ge h_2(y)$$

Denote

$\theta = [u; v], u = [x, y, z, w], v = [\tau_1, \tau_2]$

$$\Theta = \big\{\theta = [u; v] : x, y \in X, z \in Z$$
$$\|w\| \le 1, \tau_1 \ge h_1(x), \tau_2 \ge h_2(y)\big\}$$

Variational Inequality $VI(\Theta, F)$

$$\langle F(\theta), \theta - \theta_*\rangle \ge 0, \forall \theta \in \Theta$$

where

$$F(\theta) = [F_u(u); F_v]$$

$$F_u(u) = \begin{bmatrix} \nabla_x \Phi(x,z) + \rho w \\ -\rho w \\ -\nabla_z \Phi(x,z) \\ \rho(y-x) \end{bmatrix}$$

$$F_v = [1; 1]$$

Semi-structured variational inequality

$$VI(\Theta, F): \quad \langle F(\theta), \theta - \theta_* \rangle \geq 0, \forall \theta \in \Theta$$

- Decomposition of $F(\theta = [u, v]) = [F_u(u); F_v]$:

$$\|F_u(u) - F_u(u')\| \leq L\|u - u'\| \text{ and } F_v \text{ is constant}$$

- Decomposition of $\Theta = \Theta_1 \times \Theta_2$:

$$\theta = [\theta_1 = [u_1, v_1]; \theta_2 = [u_2, v_2]]$$

- "Computation-friendly" proximal mapping on Θ_1:

$$\text{Prox}(\eta_1, \alpha) := \min_{\theta_1 \in \Theta_1} \{\omega_1(u_1) + \langle \eta_1, u_1 \rangle + \alpha \langle F_{v_1}, v_1 \rangle\}.$$

for some distance generating function $\omega_1(\cdot)$.

- "Computation-friendly" linear minimization on Θ_2:

$$\text{LMO}(\eta_2, \alpha) := \min_{\theta_2 \in \Theta_2} \{\langle \eta_2, u_2 \rangle + \alpha \langle F_{v_2}, v_2 \rangle\}.$$

Composite Mirror Prox with inexact prox-mappings

$$VI(\Theta, F): \quad \langle F(\theta), \theta - \theta_* \rangle \geq 0, \forall \theta \in \Theta$$
$$F(\theta) = [F_u(u); F_v]$$

- ϵ-prox-mapping on entire Θ: let $\omega(u)$ be the d.s.g. on u-space, for any $\xi = [\eta, \zeta]$ and $\theta_0 = [u_0, v_0] \in \Theta$,

$$P^{\epsilon}_{\theta_0}(\xi) = \{[\widehat{u}, \widehat{v}] \in \Theta : \langle \eta + \omega'(\widehat{u}) - \omega'(u_0), \widehat{u} - u \rangle + \langle \zeta, \widehat{v} - v \rangle \leq \epsilon, \ \forall [u; v] \in \Theta\}$$

Inexact Composite Mirror Prox

$$\begin{array}{lrcl} (a) & \theta^1 := [u^1; v^1] & \in & \Theta; \\ (b) & \widehat{\theta}^t := [\widehat{u}^t; \widehat{v}^t] & \in & P^{\epsilon_t}_{\theta^t}(\gamma_t [F_u(u^t); F_v]) \\ (c) & \theta^{t+1} := [u^{t+1}; v^{t+1}] & \in & P^{\epsilon_t}_{\theta^t}(\gamma_t [F_u(\widehat{u}^t); F_v]) \end{array}$$

- When $\epsilon_t = 0$, reduces to the composite Mirror Prox [HJN'13] with exact prox-mapping

$$P_{\theta^t}(\gamma_t [F_u(u^t); F_v]) = \operatorname*{Argmin}_{\theta = [u;v] \in \Theta} \left\{ D_\omega(u, u^t) + \langle \gamma_t F_u(u^t), u \rangle + \langle \gamma_t F_v, v \rangle \right\}.$$

Composite Mirror Prox with inexact prox-mappings

Theorem [HH'15]

Let $\bar{\theta}_T = \left[\sum_{t=1}^T \gamma_t\right]^{-1} \sum_{t=1}^T \gamma_t \widehat{\theta}^t$ be generated by the inexact composite Mirror Prox algorithm by setting $0 < \gamma_t \leq \frac{1}{L}$ and inexactness $\epsilon_t \geq 0$. Let Ω be the bound for prox function. Then

$$\epsilon_{\mathrm{VI}}(\bar{\theta}_T \big| \Theta, F) := \sup_{\theta \in \Theta} \langle F(\theta), \bar{\theta}_T - \theta \rangle \leq \frac{\Omega + 2\sum_{t=1}^T \epsilon_t}{\sum_{t=1}^T \gamma_t}.$$

Immediate observations

- When $\epsilon_t \equiv 0$, one has at least $\epsilon_{\mathrm{VI}}(\bar{\theta}_T \big| \Theta, F) \leq \frac{\Omega L}{T}$.
- When $\epsilon_t \equiv \frac{\Omega}{2T}$, one has at least $\epsilon_{\mathrm{VI}}(\bar{\theta}_T \big| \Theta, F) \leq \frac{2\Omega L}{T}$.
- When $\epsilon_t = \frac{\Omega}{2t}$, one has at least $\epsilon_{\mathrm{VI}}(\bar{\theta}_T \big| \Theta, F) \leq \frac{\Omega L(1+\ln T)}{T}$.

Inexact prox-mapping via LMO routines

- Recall the subproblem of prox-mapping in composite Mirror Prox

$$P_{\theta^t}(\gamma_t[F_u(u^t); F_v]) = \operatorname*{Argmin}_{\theta=[u;v]\in\Theta} \{D_\omega(u, u^t) + \langle \gamma_t F_u(u^t), u\rangle + \langle \gamma_t F_v, v\rangle\}.$$

  reduces to solve the smooth semi-linear problem:

$$\min_{\theta=[u,v]\in\Theta} \left\{\phi^+(u, v) = \phi(u) + \langle c, v\rangle\right\}.$$

- Assume we have at disposal only the linear minimization oracles

$$\mathrm{LMO}(\eta) := \min_{\theta=[u,v]\in\Theta} \{\langle \eta, u\rangle + \langle c, v\rangle\}.$$

Composite Conditional Gradient

$$\min_{\theta=[u,v]\in\Theta} \left\{\phi^+(u,v) = \phi(u) + \langle c, v\rangle\right\}.$$

Composite Conditional Gradient $CCG(\Theta, \phi, c; \epsilon)$

(a) $\theta^1 := [u^1; v^1] \in \Theta$;

(b) $\widehat{\theta}^t := [\widehat{u}^t; \widehat{v}^t] = LMO(\nabla\phi(u^t))$

compute $\delta_t = \langle \nabla\phi^+(\theta^t), \theta^t - \widehat{\theta}^t\rangle$, return if $\delta_t \le \epsilon$

(c) $\theta^{t+1} := [u^{t+1}; v^{t+1}]$ s.t. $\phi^+(\theta^{t+1}) \le \phi^+(\theta^t + \frac{2}{t+1}(\widehat{\theta}^t - \theta^t))$

- Choices of θ^{t+1} in (c):
  - Simplest: $\theta^{t+1} = \theta^t + \gamma_t(\widehat{\theta}^t - \theta^t)$ with $\gamma_t = 2/(t+1)$.
  - Line search: $\theta^{t+1} = \operatorname{argmin}_{0\le\gamma\le1} \phi^+(\gamma\widehat{\theta}^t + (1-\gamma)\theta^t)$.
  - Fully-corretive search: $\theta_{t+1} = \operatorname{argmin}_{\theta\in\mathrm{conv}(\theta^1,\ldots,\theta^t)} \phi^+(\theta)$
- Use the dual gap δ_t to check the condition for ϵ-prox mapping.

Composite Conditional Gradient

Theorem [HJN'13, HH'15]

Assume convex function ϕ is (κ, L_0)-smooth for some $1 < \kappa \leq 2$,

$$\phi(u') \leq \phi(u) + \langle \nabla\phi(u), u' - u\rangle + \frac{L_0}{\kappa}\|u' - u\|^\kappa \ \forall u, u';$$

Let D the $\|\cdot\|$-diameter of $P_u\Theta$. Then

$$\phi^+(x^t) - \min_{x\in X}\phi^+(x) \leq \frac{2L_0D^\kappa}{\kappa(3-\kappa)}\left(\frac{2}{t+1}\right)^{\kappa-1}.$$

In addition, the accuracy certificates δ_t satisfy

$$\min_{1\leq s\leq t}\ \delta_s \leq O(1)L_0D^\kappa\left(\frac{2}{t+1}\right)^{\kappa-1}.$$

- When $\phi(u) = \frac{1}{2}\|u - u_0\|_2^2$, the worst complexity bound for LMO calls is $\mathcal{O}(1/\epsilon)$.

Returning to semi-structured variational inequality

$$VI(\Theta, F): \quad \langle F(\theta), \theta - \theta_* \rangle \geq 0, \forall \theta \in \Theta$$

- Decomposition of $F(\theta = [u, v]) = [F_u(u); F_v]$:

$$\|F_u(u) - F_u(u')\| \leq L\|u - u'\| \text{ and } F_v \text{ is constant}$$

- Decomposition of $\Theta = \Theta_1 \times \Theta_2$:

$$\theta = [\theta_1 = [u_1, v_1]; \theta_2 = [u_2, v_2]]$$

- "Computation-friendly" proximal mapping on Θ_1:

$$\text{Prox}(\eta_1, \alpha) := \min_{\theta_1 \in \Theta_1} \{\omega_1(u_1) + \langle \eta_1, u_1 \rangle + \alpha \langle F_{v_1}, v_1 \rangle\}.$$

for some distance generating function $\omega_1(\cdot)$.

- "Computation-friendly" linear minimization on Θ_2:

$$\text{LMO}(\eta_2, \alpha) := \min_{\theta_2 \in \Theta_2} \{\langle \eta_2, u_2 \rangle + \alpha \langle F_{v_2}, v_2 \rangle\}.$$

Working horse: Semi-Proximal Mirror-Prox

Semi-Proximal Mirror-Prox

Input: stepsizes $\gamma_t > 0$, accuracies $\epsilon_t \geq 0$, $t = 1, 2, \ldots$
[1] Initialize $\theta^1 = [\theta_1^1; \theta_2^1] \in \Theta$, where $\theta_1^1 = [u_1^1; v_1^1]; \theta_2^1 = [u_2^1, ; v_2^1]$.
for $t = 1, 2, \ldots, T$ **do**
[2] Compute $\widehat{\theta}^t = [\widehat{\theta}_1^t; \widehat{\theta}_2^t]$ that

$$\widehat{\theta}_1^t := [\widehat{u}_1^t; \widehat{v}_1^t] = \mathrm{Prox}(\gamma_t F_{u_1}(u_1^t) - \omega_1'(u_1^t), \gamma_t)$$
$$\widehat{\theta}_2^t := [\widehat{u}_2^t; \widehat{v}_2^t] = \mathrm{CCG}(X_2, \omega_2(\cdot) + \langle \gamma_t F_{u_2}(u_2^t) - \omega_2'(u_2^t), \cdot \rangle, \gamma_t F_{v_2}; \epsilon_t)$$

[3] Compute $\theta^{t+1} = [\theta_1^{t+1}; \theta_2^{t+1}]$ that

$$\theta_1^{t+1} := [u_1^{t+1}; v_1^{t+1}] = \mathrm{Prox}(\gamma_t F_{u_1}(\widehat{u}_1^t) - \omega_1'(u_1^t), \gamma_t)$$
$$\theta_2^{t+1} := [u_2^{t+1}; v_2^{t+1}] = \mathrm{CCG}(X_2, \omega_2(\cdot) + \langle \gamma_t F_{u_2}(\widehat{u}_2^t) - \omega_2'(u_2^t), \cdot \rangle, \gamma_t F_{v_2}; \epsilon_t)$$

end for
Output: $\overline{\theta}_T := [\bar{u}_T; \bar{v}_T] = \left(\sum_{t=1}^T \gamma_t\right)^{-1} \sum_{t=1}^T \gamma_t \widehat{\theta}^t$

Interpretating "semi-proximal"

- When Θ_2 is singleton, we get *full-proximal setup*, the algorithm reduces to composite Mirror Prox.
- When Θ_1 is singleton, we get *full-LMO setup*, the algorithm serves as a viable alternative to existing LMO-based algorithms via smoothing technique or dual approach for solving nonsmooth composite minimization.
- In the gray zone in between, we get *semi-proximal setup*, the algorithm is the first of its kind to solve semi-structured variational inequality, which covers the aforementioned problem of interest.

Complexity

Theorem

For the outlined algorithm to return an ϵ-solution to the $VI(\Theta, F)$, the total number of Mirror Prox steps required does not exceed

$$\mathrm{T} = O\left(\frac{L\Omega}{\epsilon}\right)$$

and the total number LMO calls required does not exceed

$$\mathcal{N} = O(1)\left(\frac{L_0 L^\kappa D^\kappa}{\epsilon^\kappa}\right)^{\frac{1}{\kappa-1}} \Omega.$$

In particular, if we use $\omega_2(\cdot) = \frac{1}{2}\|x_2\|^2$, then $\mathcal{N} = O(1)L^2 D^2 \Omega/\epsilon^2$.

Immediate observation

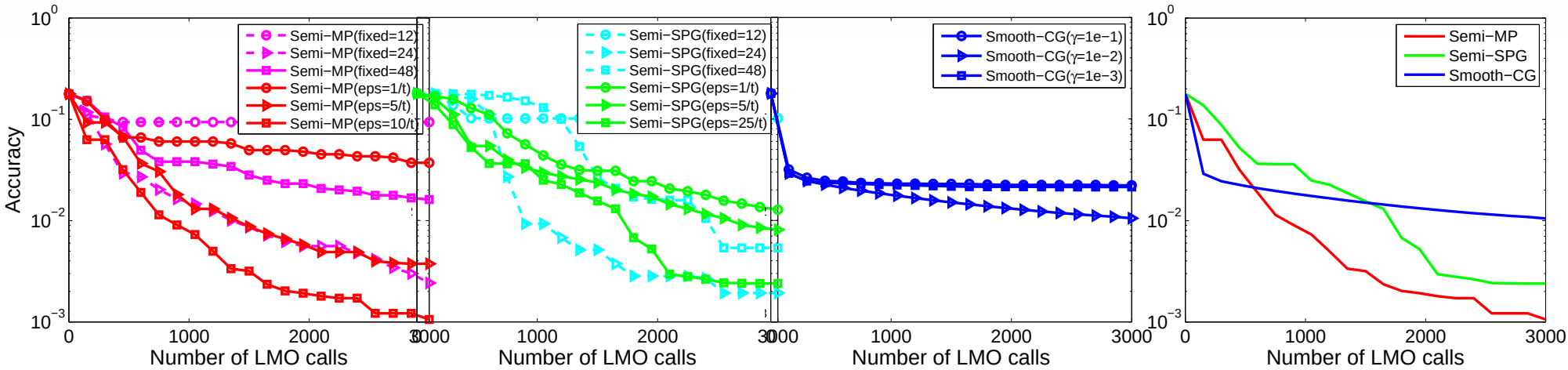
- optimal complexity bound $\mathcal{O}(1/\epsilon)$ in first-order oracles
- optimal complexity bound $\mathcal{O}(1/\epsilon^2)$ in linear minimization oracles

Numerical expriments

- Several types of problems
  - Matrix completion
  - Robust collaborative filtering
  - Link prediction
- Several alternative algorithms
  - **Semi-MP:** Semi-Proximal Mirror-Prox
  - **Smooth-CG:** Smoothing conditional gradient [PHM'14]
  - **Semi-SPG:** Smoothing proximal gradient with inexact prox-mappings via LMO routines
  - **Semi-LPADMM:** Linearized preconditioned ADMM with inexact prox-mappings via LMO routines
- Several choices of inexactness
  - Use decaying error $\epsilon_t = \frac{c}{t}$
  - Use fixed inner CG steps

Matrix completion: ℓ_2-fit + nuclear norm

$$\min_{x \in \mathbf{R}^{m \times n}} \|P_\Omega x - b\|_2 + \lambda \|x\|_{\text{nuc}}$$

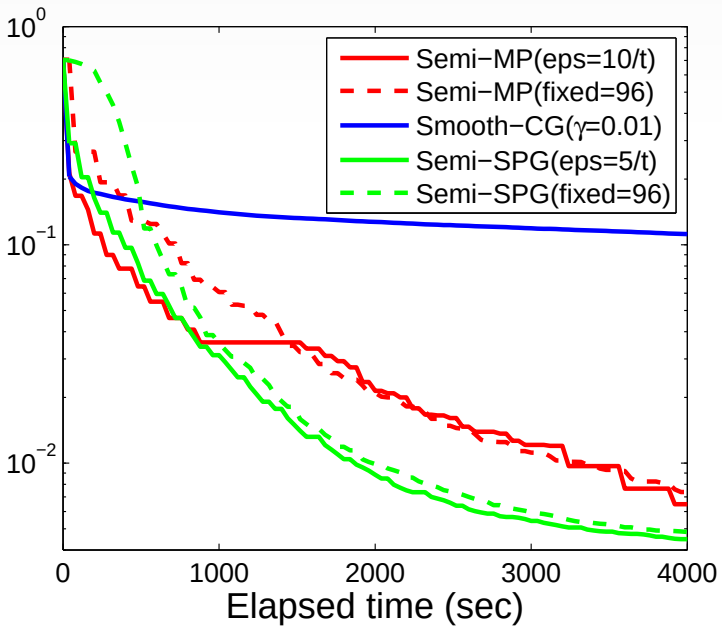


(a)Semi-MP (b)Semi-SPG (c)Smooth-CG (d)best of three

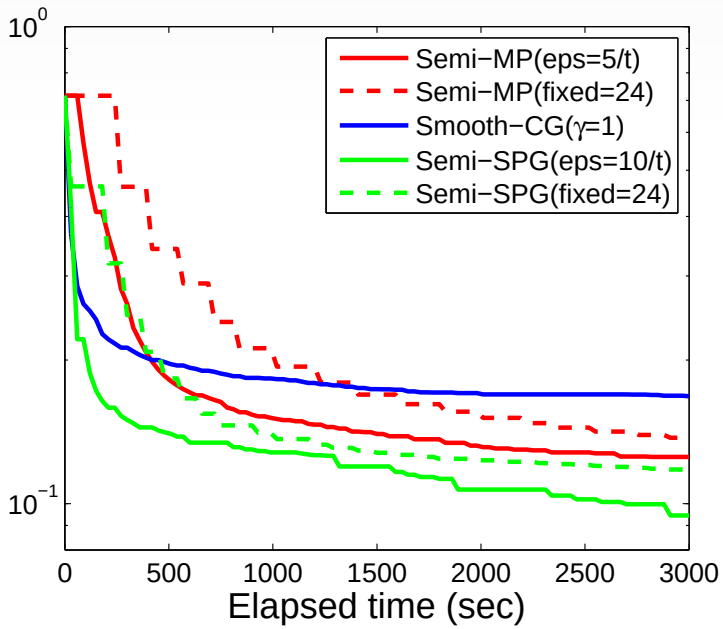
Synthetic dataset(1024×1024): optimality gap vs the LMO calls.

Robust collaborative filtering: ℓ_1-loss + nuclear norm

$$\min_{x \in \mathbf{R}^{m \times n}} \frac{1}{|E|} \sum_{(i,j) \in E} |x_{ij} - b_{ij}| + \lambda \|x\|_{\text{nuc}}$$



(a) MovieLens 100K(943×1682)

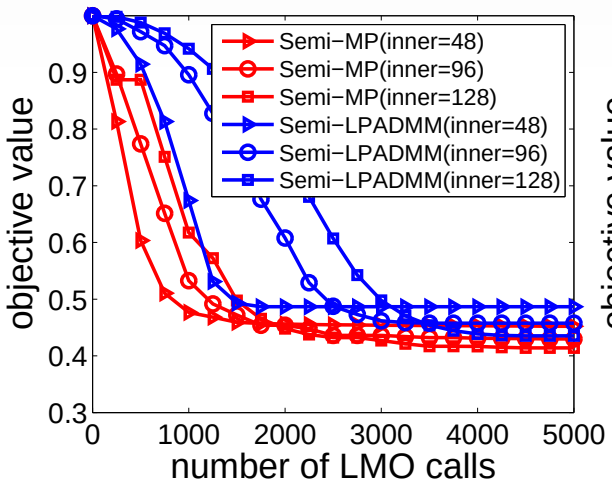


(b) MovieLens 1M(3952×6040)

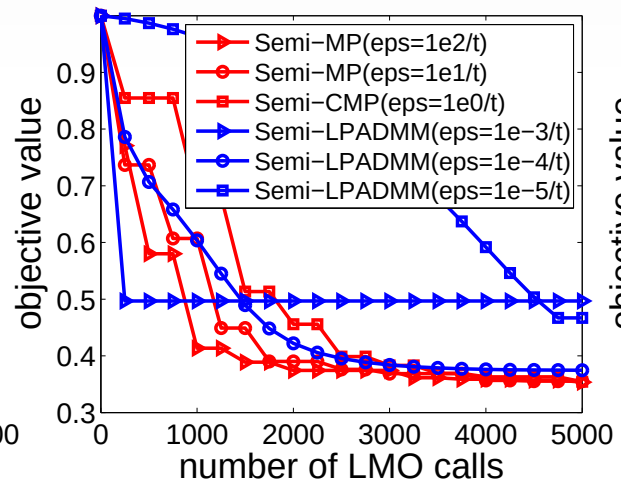
MovieLens dataset: objective function vs elapsed time.

Link prediction: hinge loss + ℓ_1-norm + nuclear norm

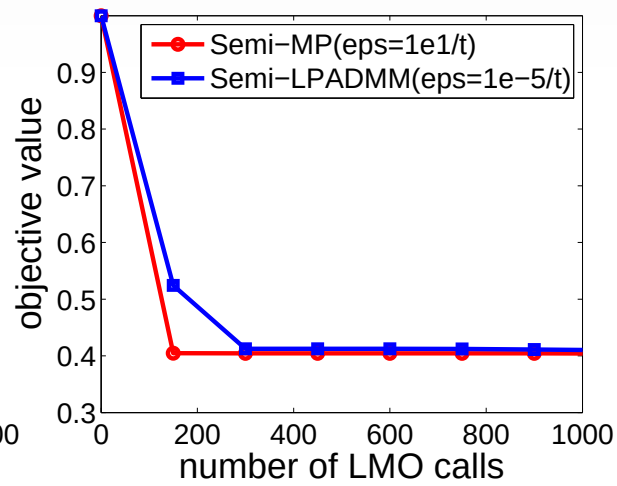
$$\min_{x \in \mathbf{R}^{m \times n}} \frac{1}{|E|} \sum_{(i,j) \in E} \max(1 - b_{ij} x_{ij}, 0) + \lambda_1 \|x\|_1 + \lambda_2 \|x\|_{\text{nuc}}$$



(a)Wikivote(1024 nodes) fixed inner steps



(b)Wikivote(1024 nodes) decaying $\epsilon_t = c/t$



(c)Wikivote(7118 nodes) decaying $\epsilon_t = c/t$

Link prediction on Wikivote: objective function value against the LMO calls.

Conclusion and perspectives

- We present a new algorithm, Semi-Proximal Mirror-Prox , that solves a broad class of composite minimization and leverages
  - saddle point representation of f
  - proximal operatiors of h_1
  - linear minimization oracles of h_2
- Theoretically, the algorithm achieves
  - optimal complexity bound $\mathcal{O}(1/\epsilon)$ in first-order oracles
  - optimal complexity bound $\mathcal{O}(1/\epsilon^2)$ in linear minimization oracles
- Practically, the algorithm shows advantages in comparison to competing methods in various large-scale applications.
- The algorithm is readily extensible to online/stochastic settings.

Thank you!

References

- [HH'15] *Semi-proximal Mirror-Prox for nonsmooth composite minimization*, N. He and Z. Harchaoui, 2015.
- [HJN'15] *Mirror Prox algorithm for multi-term composite minimization and semi-separable problems*, N. He, A. Juditsky, and A. Nemirovksi, 2015.
- [HJN'13] *Conditional gradient algorithms for norm-regularized smooth convex optimization*, Z. Harchaoui, A. Juditsky, and A. Nemirovksi, 2013.
- [Lan'13] *The complexity of large-scale convex programming under a linear optimization oracle*, G. Lan, 2013.
- [LZ'14] *Conditional gradient sliding for convex optimization*, G. Lan and Y. Zhou, 2014.
- [PHM'14] *A smoothing approach for composite conditional gradient with nonsmooth loss*, F. Pierucci, Z. Harchaoui, and J. Malick, 2014.