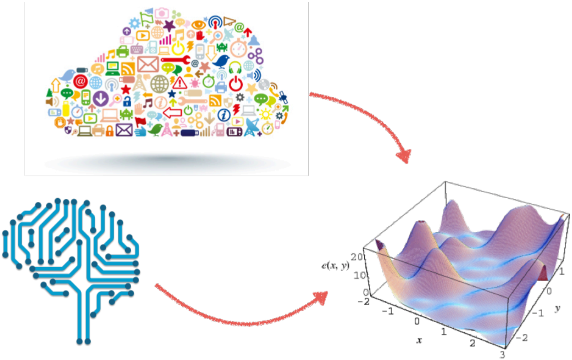
A Projection-free Decentralized Algorithm for Non-convex Optimization

Hoi-To Wai[‡], Anna Scaglione[‡], Jean Lafond[†] and Eric Moulines[#]

[‡] School of ECEE, Arizona State University, USA.
[†] Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, France.
[#] CMAP, Ecole Polytechnique, Palaiseau, France.

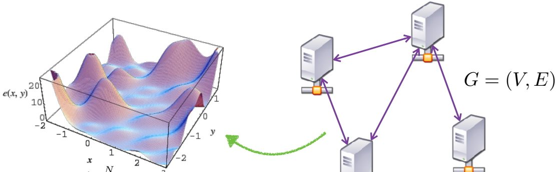
Acknowledgement: Direction Générale de l'Armement and the labex LMH (ANR-11-LABX-0056-LMH), NSF CCF-1011811.

Motivation



- Big data, machine learning $\Longrightarrow$ non-convex, high-dim. optimization.
- Decentralized/multi-agent opt. exploits the collective computation power and allows sharing of data among the agents.

Problem Setup



Ensemble loss: $F(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{\theta})$

Local loss function: $f_i(\boldsymbol{\theta})$

- $G = (V, E)$ – connected graph with N agents.
- We consider:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} F(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{\theta}) \text{ s.t. } \boldsymbol{\theta} \in \mathcal{C} \ . \qquad \text{(P1)}$$

  - $f_i : \mathbb{R}^d \to \mathbb{R}$ – **smooth** loss function of agent i $\sim$ data owned by agent i (possibly **non-convex**).
  - $\mathcal{C} \subseteq \mathbb{R}^d$ – **convex** and **compact** constraint ($\sim$regularization).
- **Goal**: tackle (P1) with agents **only** communicating on G.

Prior Works

- **Proximal/Projected gradient (PG)** — [RNV12, JXM14, SLWY15]
    - works for time varying graph and asynchronous algorithm.
    - most analysis only work for convex problems except for [BJ13, GL15].
- **Primal-dual approach** — [CNS14, Hon16]
    - able to handle more complicated constraints.
    - requires convexity except for [Hon16].
- **Projection-free/Frank-Wolfe (FW)** — [Jag13]
    - efficient for high dimensional problems which are costly to run PG on.
    - centralized algorithm for convex opt. except for [LJ16, RSPS16].
    - **This work**: decentralized FW & its convergence for non-convex opt.
- **Others** —
    - second order method [LS13].
    - decomposition by block coordinate descent [LS16].
    - convergence rates are not analyzed in these works.

Prior Works

- **Proximal/Projected gradient (PG)** — [RNV12, JXM14, SLWY15]
  - works for time varying graph and asynchronous algorithm.
  - most analysis only work for convex problems except for [BJ13, GL15].
- **Primal-dual approach** — [CNS14, Hon16]
  - able to handle more complicated constraints.
  - requires convexity except for [Hon16].
- **Projection-free/Frank-Wolfe (FW)** — [Jag13]
  - efficient for high dimensional problems which are costly to run PG on.
  - centralized algorithm for convex opt. except for [LJ16, RSPS16].
  - **This work**: decentralized FW & its convergence for non-convex opt.
- **Others** —
  - second order method [LS13].
  - decomposition by block coordinate descent [LS16].
  - convergence rates are not analyzed in these works.

Curse of Dimensionality — Why projection-free?

- Decentralized PG algorithm [RNV12] — for all $i \in [N]$ and $\gamma_t \in (0, 1]$,

$$\bar{\boldsymbol{\theta}}_i^t \leftarrow \underbrace{\text{LocalAvg}(\{\boldsymbol{\theta}_j^t\}_{j \in \mathcal{N}_i})}_{\text{e.g., by gossiping: } \sum_{j=1}^N W_{ij} \boldsymbol{\theta}_j^t} \;,\; \boldsymbol{\theta}_i^{t+1} \leftarrow \underbrace{\mathcal{P}_{\mathcal{C}}}_{\text{Projection Operator}} \left(\bar{\boldsymbol{\theta}}_i^t - \gamma_t \nabla f_i(\bar{\boldsymbol{\theta}}_i^t) \right) .$$

- Computing $\mathcal{P}_{\mathcal{C}} : \mathbb{R}^d \to \mathbb{R}^d$ may require substantial complexity, e.g.,
  - If $\mathcal{C}$ is the trace-norm ball for $m_1 \times m_2$ matrices with radius r, then

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{\theta}) = \boldsymbol{U} \boldsymbol{\Sigma}^+ \boldsymbol{V}^\top, \;\; \boldsymbol{\Sigma}^+ = \text{Diag}(\max\{0, \boldsymbol{\sigma}(\boldsymbol{\theta}) - \lambda^\star(r) \mathbf{1}\}) \;, \tag{1}$$

    where $\boldsymbol{U}, \boldsymbol{V}$ are left/right singular vectors of $\boldsymbol{\theta} \in \mathbb{R}^{m_1 \times m_2}$, $\lambda^\star(r) \geq 0$ is a Lagrangian multiplier and $\boldsymbol{\sigma}(\boldsymbol{\theta})$ are the singular values of $\boldsymbol{\theta}$.
  - requires the **Full SVD** $\Longrightarrow \mathcal{O}((m_1 \wedge m_2)^3)$ per iteration & per agent.

- **Frank-Wolfe** (FW, a.k.a. projection-free) optimization reduces per iteration complexity to $\mathcal{O}(m_1 \wedge m_2)$ for the example above.

Agenda

The centralized FW Algorithm

- The (centralized) FW algorithm — $\gamma_t \in (0,1]$ is a step size,

$$\boldsymbol{\theta}^{t+1} \leftarrow \boxed{(1-\gamma_t)\boldsymbol{\theta}^t + \gamma_t \boldsymbol{a}^t} \quad \text{where} \quad \boldsymbol{a}^t = \boxed{\arg\min_{\boldsymbol{a}\in\mathcal{C}} \ \langle \boldsymbol{a}, \nabla F(\boldsymbol{\theta}^t)\rangle} \ . \qquad (2)$$

- Update direction $\boldsymbol{a}^t \approx$ most correlated vector in $\mathcal{C}$ with *negative gradient*.
- Param. update: $\boldsymbol{\theta}^{t+1}$ is a convex combination between $\boldsymbol{a}^t$ and $\boldsymbol{\theta}^t$.

Convergence of (centralized) FW algorithm

- If $F(\boldsymbol{\theta})$ is convex and smooth, and $\gamma_t = 1/t$, then $F(\boldsymbol{\theta}^t) - F(\boldsymbol{\theta}^\star) = \mathcal{O}(1/t)$ [FW56], where $\boldsymbol{\theta}^\star$ is an optimal solution to (P1).
- If $F(\boldsymbol{\theta})$ is non-convex and smooth, and $\gamma_t = t^{-\alpha}$ with $\alpha > 0.5$, then the limit points of the sequence $\{\boldsymbol{\theta}^t\}_{t=1}^{\infty}$ are *stationary points* of (P1) [WLSM16].

Advantage of FW over PG

- The (centralized) FW algorithm — $\gamma_t \in (0, 1]$ is a step size,

$$\boldsymbol{\theta}^{t+1} \leftarrow \boxed{(1-\gamma_t)\boldsymbol{\theta}^t + \gamma_t \boldsymbol{a}^t} \text{ where } \boldsymbol{a}^t = \boxed{\arg\min_{\boldsymbol{a} \in \mathcal{C}} \langle \boldsymbol{a}, \nabla F(\boldsymbol{\theta}^t) \rangle} .$$

- Requires only a Linear Optimization (LO)
  - This LO step 'replaces' the projection operation in PG.
  - If $\mathcal{C}$ is the trace-norm ball for $m_1 \times m_2$ matrices with radius r, then

$$\boldsymbol{a}^t = -r \cdot \boldsymbol{u}_1 \boldsymbol{v}_1^\top , \tag{3}$$

    where $\boldsymbol{u}_1, \boldsymbol{v}_1$ are the top left/right singular vectors.
  - requires only **Principal Component** $\Longrightarrow \mathcal{O}(m_1 \wedge m_2)$ per iteration.
  - recall that the PG method has $\mathcal{O}((m_1 \wedge m_2)^3)$ per iteration.

A Perturbed Frank-Wolfe Algorithm

- Let $\bar{\boldsymbol{\theta}}^t := (1/N)\sum_{j=1}^N \boldsymbol{\theta}_j^t$. Consider a perturbed FW algorithm —

$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boxed{(1-\gamma_t)\bar{\boldsymbol{\theta}}_i^t + \gamma_t \boldsymbol{a}_i^t} \quad \text{where} \quad \boldsymbol{a}_i^t \leftarrow \boxed{\arg\min_{\boldsymbol{a}\in\mathcal{C}} \ \langle \boldsymbol{a}, \overline{\nabla_i^t F}\rangle} \ , \tag{4}$$

where $\bar{\boldsymbol{\theta}}_i^t$ and $\overline{\nabla_i^t F}$ are *perturbed* version of $\bar{\boldsymbol{\theta}}^t$ and $\nabla F(\bar{\boldsymbol{\theta}}^t)$:

$$\bar{\boldsymbol{\theta}}_i^t \approx \boxed{(1/N)\sum_{j=1}^N \boldsymbol{\theta}_j^t} \quad \overline{\nabla_i^t F} \approx \boxed{(1/N)\sum_{j=1}^N \nabla f_j(\bar{\boldsymbol{\theta}}_j^t)} \approx \nabla F(\bar{\boldsymbol{\theta}}^t) \ .$$

- *Special case:* when both approximations are **exact**
  - Eq. (4) is equivalent to a centralized FW on the iterates $\{\bar{\boldsymbol{\theta}}^t\}_{t=1}^\infty$.
- The iterates $\{\bar{\boldsymbol{\theta}}_i^t\}_{t\geq 1} \approx$ running *perturbed* FW on $\{\bar{\boldsymbol{\theta}}^t\}_{t\geq 1}$.

Convergence Result (1)

- Assuming that the approximation accuracy improves with t,

H1 : $\|\bar{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}^t\| \leq C_g t^{-\alpha}$ and $\|\overline{\nabla_i^t F} - N^{-1} \sum_{j=1}^N \nabla f_j(\bar{\boldsymbol{\theta}}_j^t)\| \leq C_p t^{-\alpha}$.

Theorem 1 (Convergence of perturbed FW)

Suppose that F is L-smooth, G-Lipschitz and **H1** holds. With $\gamma_t = t^{-\alpha}$, $\alpha \in [0.5, 1)$:

$$\begin{aligned} &\min_{t \in [T/2+1, T]} \max_{\boldsymbol{\theta} \in \mathcal{C}} \langle \nabla F(\bar{\boldsymbol{\theta}}^t), \bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta} \rangle \leq \\ &\qquad \frac{1}{T^{1-\alpha}} \cdot \frac{1-\alpha}{(1-(2/3)^{1-\alpha})} \cdot \Big(G\bar{\rho} + (L\bar{\rho}^2/2 + 2\bar{\rho}(C_g + LC_p)) \log 2 \Big) , \end{aligned} \tag{5}$$

for all $T \geq 6$, where $\bar{\rho} := \sup_{\boldsymbol{\theta}', \boldsymbol{\theta} \in \mathcal{C},\ \boldsymbol{\theta} \neq \boldsymbol{\theta}'} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$.

- In [WLSM16], we also show that the limit points of the sequence $\{\bar{\boldsymbol{\theta}}^t\}_{t=1}^{\infty}$ are stationary points of (P1) if $\alpha > 0.5$.

Proof Idea

Define $g_t := \max_{\boldsymbol{\theta} \in \mathcal{C}} \langle \nabla F(\bar{\boldsymbol{\theta}}^t), \bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta} \rangle$. With L-smoothness of F, we have

$$F(\bar{\boldsymbol{\theta}}^{t+1}) \leq F(\bar{\boldsymbol{\theta}}_t) - \gamma_t g_t + 2t^{-\alpha} \bar{\rho} \cdot (C_g t^{-\alpha} + L \cdot C_p t^{-\alpha}) + t^{-2\alpha} \frac{L\bar{\rho}^2}{2} \; . \tag{6}$$

This implies

$$\sum_{t=T/2+1}^{T} \gamma_t g_t \leq \sum_{t=T/2+1}^{T} \Bigg(\underbrace{F(\bar{\boldsymbol{\theta}}^t) - F(\bar{\boldsymbol{\theta}}^{t+1})}_{\text{terms can be cancelled} \implies \text{bounded by } G\bar{\rho}.} + \mathcal{O}(t^{-2\alpha}) \Bigg) \; . \tag{7}$$

- By definition, we have $g_t \geq 0$ for all t.
- Left hand side is **lower bounded** by $\Omega(T^{1-\alpha}) \cdot \min_{t \in [T/2+1, T]} g_t$.
- Right hand side is **upper bounded** by $\mathcal{O}(1)$.

Convergence Result (2)

- Under **H1**, for $\alpha \in [0.5, 1)$, the perturbed FW algorithm yields

$$\min_{t \in [T/2+1, T]} \underbrace{\max_{\boldsymbol{\theta} \in \mathcal{C}} \langle \nabla F(\bar{\boldsymbol{\theta}}^t), \bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta} \rangle}_{:=\textbf{FW gap} \text{ (a.k.a. 'duality' gap)}} = \mathcal{O}(1/T^{1-\alpha}), \; \forall \, T \geq 6 \; ,$$

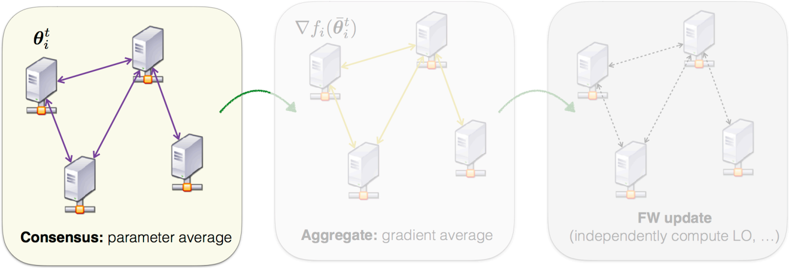
- If the FW gap becomes zero, then

$$\langle \nabla F(\bar{\boldsymbol{\theta}}^t), \bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta} \rangle \leq 0, \; \forall \, \boldsymbol{\theta} \in \mathcal{C} \; .$$

- $\Longrightarrow$ The parameter $\bar{\boldsymbol{\theta}}^t$ in the above is a stationary point to (P1).
- Fastest rate is when $\alpha = 0.5$, giving us $\mathcal{O}(1/\sqrt{T})$.
- **Remaining task** — how do we satisfy **H1**?
- Needs approximate averages $\bar{\boldsymbol{\theta}}^t$, $\frac{1}{N} \sum_{i=1}^N f_i(\bar{\boldsymbol{\theta}}_i^t) \Longrightarrow$ Gossiping!

Decentralized FW (DeFW) algorithm via Gossiping

- $\boldsymbol{W} \in \mathbb{R}_+^{N \times N}$ is *doubly stochastic* and $W_{ij} = 0$ iff $ij \notin E$.
- Decentralized algorithm that relies on *in-network* computation:

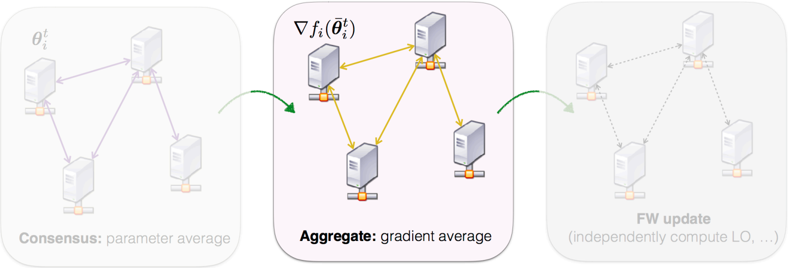


Consensus Step: (to get $\bar{\boldsymbol{\theta}}_i^t$ with H1, i.e., $\|\bar{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}^t\| = \mathcal{O}(t^{-\alpha})$)

$$\bar{\boldsymbol{\theta}}_i^{t,0} \leftarrow \boldsymbol{\theta}_i^t, \quad \text{repeat } L_t \text{ times } \left(\bar{\boldsymbol{\theta}}_i^{t,\ell+1} \leftarrow \textstyle\sum_{j=1}^N W_{ij} \bar{\boldsymbol{\theta}}_j^{t,\ell}\right), \quad \bar{\boldsymbol{\theta}}_i^t \leftarrow \bar{\boldsymbol{\theta}}_i^{t,L_t} .$$

Decentralized FW (DeFW) algorithm via Gossiping

- $\boldsymbol{W} \in \mathbb{R}_+^{N \times N}$ is *doubly stochastic* and $W_{ij} = 0$ iff $ij \notin E$.
- Decentralized algorithm that relies on *in-network* computation:

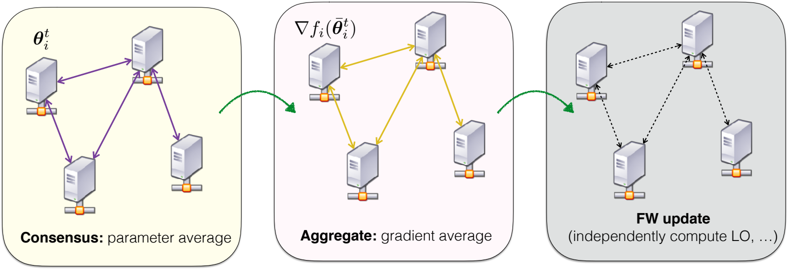


Aggregate Step: (to get $\overline{\nabla_i^t F}$ with H1)

$$\overline{\nabla_i^{t,0} F} \leftarrow \nabla f_i(\bar{\boldsymbol{\theta}}_i^t), \ \boxed{\text{repeat } L_t \text{ times } \left(\overline{\nabla_i^{t,\ell+1} F} \leftarrow \sum_{j=1}^N W_{ij} \overline{\nabla_j^{t,\ell} F}\right)}, \ \overline{\nabla_i^t F} \leftarrow \overline{\nabla_i^{t,L_t} F}.$$

Decentralized FW (DeFW) algorithm via Gossiping

- $\boldsymbol{W} \in \mathbb{R}_+^{N \times N}$ is *doubly stochastic* and $W_{ij} = 0$ iff $ij \notin E$.
- Decentralized algorithm that relies on *in-network* computation:



FW update:

$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boxed{(1 - \gamma_t)\bar{\boldsymbol{\theta}}_i^t + \gamma_t \boldsymbol{a}_i^t} \quad \text{where} \quad \boldsymbol{a}_i^t = \boxed{\arg\min_{\boldsymbol{a} \in \mathcal{C}} \ \langle \boldsymbol{a}, \overline{\nabla}_i^t F \rangle} \ .$$

DeFW Algorithm via Gossiping — Convergence

- **Gossip average consensus** (GAC) is applied to obtain $\bar{\boldsymbol{\theta}}_i^t$, $\overline{\nabla_i^t F}$.
- The GAC protocol converges **geometrically** in L_t.

Convergence of DeFW

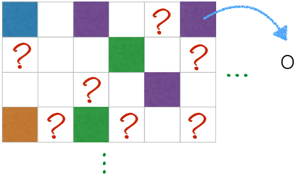
Set $L_t = (-\alpha/\log(\sigma_2(\boldsymbol{W}))) \cdot \log t$, the perturbed iterates track averages as [BGPS06]:

$$\|\overline{\nabla_i^t F} - N^{-1} \textstyle\sum_{j=1}^N \nabla f_j(\bar{\boldsymbol{\theta}}_j^t)\| = \mathcal{O}(t^{-\alpha}) \;\text{ and }\; \|\bar{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}^t\| = \mathcal{O}(t^{-\alpha}) \; .$$

As a corollary, **H1** is satisfied and Theorem 1 holds for DeFW.

- **Drawback**: number of information exchange per iteration L_t grows with t as $L_t \propto \log t$.
  - In [WLSM16], we propose an improved DeFW algorithm which only requires a **constant** no. of info. exchange $L_t = L$.
  - Key idea: using memory from the previous iteration.

Example: Sparse+Low Rank Matrix Completion (MC)



Low rank matrix $\boldsymbol{\theta}^\star$

Outliers contamination in observation

$$Y_s = [\boldsymbol{\theta}^\star]_{k_s,l_s} + Z_s$$

The noise is **"sparse"**

$$P(Z_s \neq 0) = p \ll 1$$

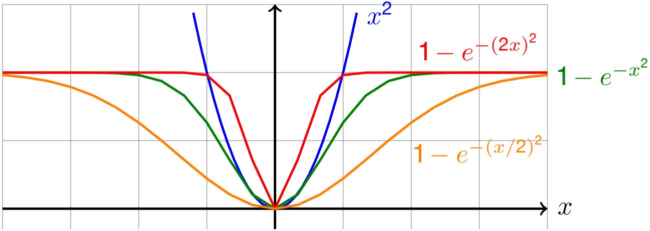
Possible scenario: data corruption in the memory, ...

- Low rank matrix $\boldsymbol{\theta}^\star \in \mathbb{R}^{m_1 \times m_2}$ is partially observed + sparse noise.
- Let $\Omega_i \subseteq [m_1] \times [m_2]$ be the observation set for agent i, we tackle:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{m_1 \times m_2}} \sum_{i=1}^{N} \sum_{(k,l)\in\Omega_i} \big(1 - \exp\big(-([\boldsymbol{\theta}]_{k,l} - Y_{k,l})^2/\sigma_i\big)\big) \text{ s.t. } \|\boldsymbol{\theta}\|_{\sigma,1} \leq r \,. \quad (8)$$

- It has a negated Gaussian loss & is a **non-convex** problem!

Example: Sparse+Low Rank Matrix Completion (MC)



- Low rank matrix $\boldsymbol{\theta}^\star \in \mathbb{R}^{m_1 \times m_2}$ is partially observed + sparse noise.
- Let $\Omega_i \subseteq [m_1] \times [m_2]$ be the observation set for agent i, we tackle:

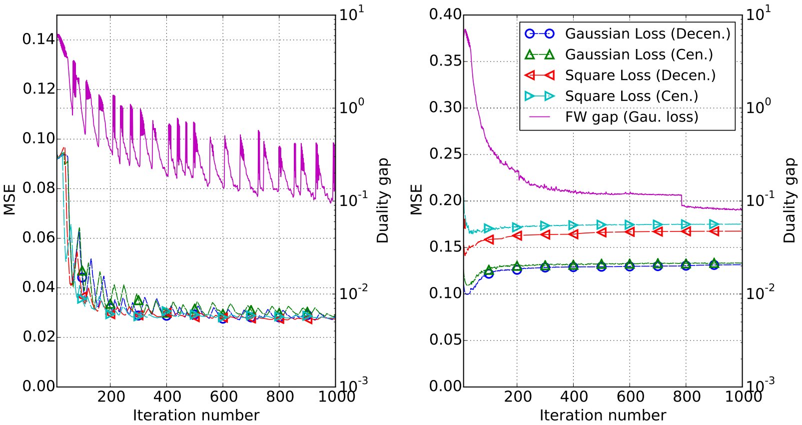
$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{m_1 \times m_2}} \sum_{i=1}^{N} \sum_{(k,l) \in \Omega_i} \big(1 - \exp\big(-([\boldsymbol{\theta}]_{k,l} - Y_{k,l})^2 / \sigma_i\big)\big) \text{ s.t. } \|\boldsymbol{\theta}\|_{\sigma,1} \leq r \ . \quad (8)$$

- It has a negated Gaussian loss & is a **non-convex** problem!

Numerical Experiment

- Simulate G as an Erdos-Renyi graph with $N = 50$ and connectivity 0.1.
- Weights on the matrix $\boldsymbol{W}$ are found with the Metropolis-Hastings rule.
- For the DeFW algorithm, we set $\gamma_t = t^{-0.75}$, $L_t = \lceil 5 + 0.75 \log t \rceil$.
- Sparse+low-rank MC problem for two datasets —
  - Synthetic dataset: $m_1 = 100$, $m_2 = 250$, $|\Omega_i| = 500$ and rank$(\boldsymbol{\theta}^\star) = 10$.
  - `movielens100k` dataset (training): $m_1 = 943$ users, $m_2 = 1682$ movies and $|\Omega_i| = 1600$ movie ratings from different users.
- Two settings tested — (i) noiseless; (ii) sparse-noise ($Z_s = p_s \tilde{Z}_s$ such that $p_s \sim B(0.1)$, $\tilde{Z}_s \sim \mathcal{N}(0, 5)$).
- Test metrics — (i) *test MSE*, *i.e.,* MSE evaluated on the testing set $[m_1] \times [m_2] \setminus \Omega$; (ii) *FW gap*, *i.e.,* $\max_{\boldsymbol{\theta} \in \mathcal{C}} \langle \nabla F(\bar{\boldsymbol{\theta}}^t), \bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta} \rangle$.

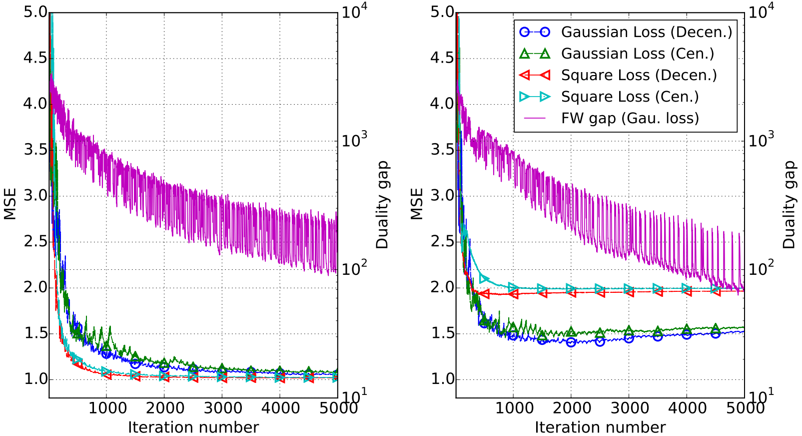
Synthetic dataset



(Left) noiseless observations; (Right) outlier-contaminated observations. Set $\sigma_i = 5$ in (8).

- DeFW algorithms converge for both convex and non-convex loss (FW gap $\to$ **0**).
- Negated Gaussian loss (non-convex) formulation is more robust to sparse noise.

Real dataset (`movielens100k`)



(Left) noiseless observations; (Right) outlier-contaminated observations. Set $\sigma_i = 5$ in (8).

- Similar observations as in the synthetic data case.
- In practice, DeFW is ~20-30 times faster than D-PG in computation time.

Conclusions

- We have proposed a decentralized, projection-free algorithm with convergence guarantee for non-convex optimization.
- The convergence results are new for projection-free algorithms in the centralized case; see recent works in [LJ16, RSPS16].
- The convergence rate is $\mathcal{O}(1/\sqrt{T}) \approx$ centralized PG analyzed in [GL15].

Future works —

- Source-privacy preserving low rank regression (submitted to ICASSP17).
- Asynchronous DeFW for time varying graph.
- Extension to primal dual optimization.

Thank you! Questions?

[BGPS06] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Trans. Inf. Theory*, 52(6):2508–2530, June 2006.

[BJ13] Pascal Bianchi and J. Jakubowicz. Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. *IEEE Trans. Autom. Control*, 58(2):391–405, Feb 2013.

[CNS14] T-H. Chang, A. Nedic, and A. Scaglione. Distributed constrained optimization by consensus-based primal-dual perturbation method. *IEEE Trans. Autom. Control*, 59(6):1524–1538, June 2014.

[EV76] Yu. M. Ermol'ev and P. I. Verchenko. A linearization method in limiting extremal problems. *Cybernetics*, 12(2):240–245, 1976.

[FW56] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Res. Logis. Quart.*, 1956.

[GL15] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, Feb 2015.

[Hon16] M. Hong. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications. *CoRR*, abs/1604.00543, Apr 2016.

[Jag13] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

[JXM14] Dusan Jakovetic, Joao Xavier, and Jose M. F. Moura. Fast distributed gradient methods. *IEEE Trans. Autom. Control*, 59(5):1131–1146, May 2014.

[LJ16] S. Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *CoRR*, July 2016.

[LS13] Xiao Li and Anna Scaglione. Convergence and applications of a gossip based gauss newton algorithm. *IEEE Trans. Signal Process.*, 61(21):5231–5246, Nov 2013.

[LS16] Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Trans. on Signal and Info. Process. over Networks*, 2(2):120–136, June 2016.

[RNV12] S. S. Ram, A. Nedic, and V. V. Veeravalli. A new class of distributed optimization algorithms : application to regression of distributed data. *Optimization Methods and Software*, (1):37–41, February 2012.

[RSPS16] S. J. Reddi, S. Sra, B. Poczos, and A. Smola. Stochastic frank-wolfe methods for nonconvex optimization. *CoRR*, July 2016.

[SLWY15] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. A Proximal Gradient Algorithm for Decentralized Composite Optimization. *IEEE Trans. on Signal Process.*, pages 1–11, 2015.

[WLSM16] H.-T. Wai, Jean Lafond, Anna Scaglione, and Eric Moulines. Decentralized projection-free optimization for convex and non-convex problems. *CoRR*, December 2016.