Primal-Dual Block Generalized Frank-Wolfe

Qi Lei*[1], Jiacheng Zhuo*[1], Constantine Caramanis[1], Inderjit S. Dhillon[1,2], and Alexandros G. Dimakis[1]

*Equal contribution [1]University of Texas at Austin. [2]Amazon

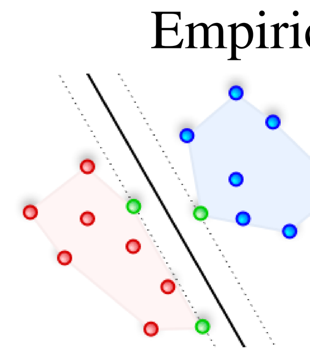
Introduction

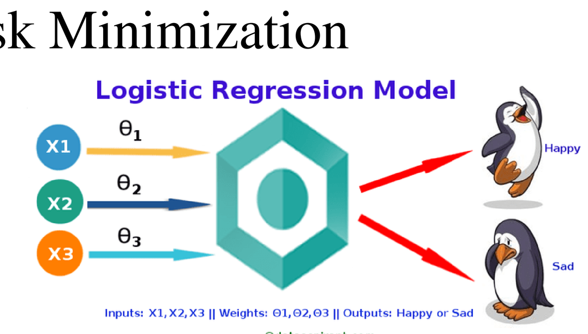
- We consider the convex-concave saddle point problem (with constraints):

$$\min_{\boldsymbol{x}\in C\subset\mathbb{R}^d}\max_{\boldsymbol{y}\in\mathbb{R}^n}\{L(\boldsymbol{x},\boldsymbol{y})=f(\boldsymbol{x})+\boldsymbol{y}^\top A\boldsymbol{x}-g(\boldsymbol{y})\}$$

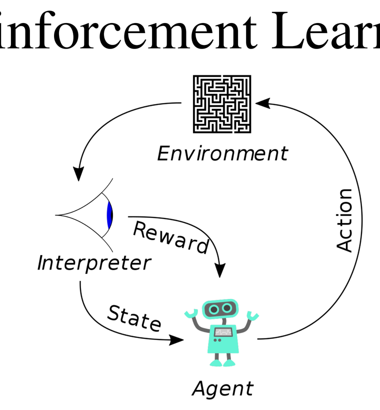
- Applications:

Empirical Risk Minimization

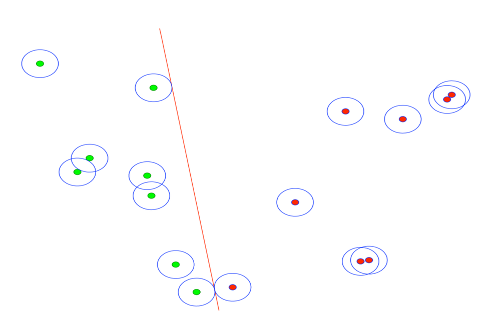




Reinforcement Learning



Robust Optimization



(Du et al. 2017) (Ben-Tal et al., 2009)

- Our focus: Primal-Dual Formulation
  - Primal form:

$$\min_{\boldsymbol{x}\in C}\left\{P(\boldsymbol{x})\equiv\frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{a}_i^\top\boldsymbol{x})+g(\boldsymbol{x})\right\},$$

C is nuclear norm ball/ℓ_1-ball

  - This form applies to Matrix Sensing, Elastic Net, Regularized SVM, Phase Retrieval, etc.
  - Its primal-dual form:

$$\min_{\boldsymbol{x}\in C}\max_{\boldsymbol{y}}\left\{\mathcal{L}(\boldsymbol{x},\boldsymbol{y})\equiv g(\boldsymbol{x})+\frac{1}{n}\langle\boldsymbol{y},A\boldsymbol{x}\rangle-\frac{1}{n}\sum_{i=1}^n f_i^*(y_i)\right\},$$

  - Its dual formulation:

$$\max_{\boldsymbol{y}}\left\{D(\boldsymbol{y}):=\min_{\boldsymbol{x}\in C}\left\{g(\boldsymbol{x})+\frac{1}{n}\langle\boldsymbol{y},A\boldsymbol{x}\rangle\right\}-\frac{1}{n}\sum_{i=1}^n f_i^*(y_i)\right\}.$$

- Our goal: design an algorithm that
  1) has time costs dependent to the structural complexity (sparsity/rank) instead of the ambient dimension,
  2) achieves linear convergence with Frank-Wolfe for strongly convex functions

Observations and Challenges On Frank-Wolfe algorithm

Lessons from constrained minimization problems:

Observations.

Frank-Wolfe conducts *partial updates:*

1. For ℓ_1 ball constraint, FW conducts **1-sparse** update
2. For nuclear norm ball constraint, FW conducts **rank-1** update

Challenges to get full benefits from FW and the partial updates.

1. FW yield **sublinear convergence** even for strongly convex problems
2. Even with partial updates, FW requires to compute the full gradient. (For big data setting, **per iteration complexity is the same with projected gradient descent**.)

Theoretical Vignette

Input: Data matrix $A\in\mathbb{R}^{n\times d}$, iteration T.
Initialize: $x_1\leftarrow 0$.
for $t=1,2,\cdots,T-1$ **do**
  Projected GD:
$$\Delta x_t\leftarrow\operatorname*{argmin}_{\|\Delta x\|_1\le\tau}\{\langle\nabla f(x_t),\Delta x\rangle+\frac{\beta}{2}\eta\|\Delta x-x_t\|_2^2\}$$
  Frank-Wolfe:
$$\Delta x_t\leftarrow\operatorname*{argmin}_{\|\Delta x\|_1\le\tau}\{\langle\nabla f(x_t),\Delta x\rangle\}$$
  Block Frank-Wolfe:
$$\Delta x_t\leftarrow\operatorname*{argmin}_{\|\Delta x\|_1\le\tau,\|\Delta x\|_0\le s}\{\langle\nabla f(x_t),\Delta x\rangle+\frac{\beta}{2}\eta\|\Delta x-x_t\|_2^2\}$$
  $x_{t+1}\leftarrow(1-\eta)x_t+\eta\Delta x_t$ (1)
end for
Output: x_T

Linear convergence of block FW: Let $h_t=f(\boldsymbol{x}_t)-f^*$.

$$h_t=f(\boldsymbol{x}_{t-1}+\eta(\Delta-\boldsymbol{x}_{t-1}))-f^*$$
$$\le h_{t-1}+\eta\langle\nabla f(\boldsymbol{x}_{t-1}),\Delta-\boldsymbol{x}_{t-1}\rangle+\frac{L}{2}\eta^2\|\Delta-\boldsymbol{x}_{t-1}\|^2$$
$$\le h_{t-1}+\eta\langle\nabla f(\boldsymbol{x}_{t-1}),\boldsymbol{x}^*-\boldsymbol{x}_{t-1}\rangle+\frac{L}{2}\eta^2\|\boldsymbol{x}^*-\boldsymbol{x}_{t-1}\|^2$$
$$\le(1-\eta+\frac{L}{\mu}\eta^2)h_{t-1}$$

Reduce Iteration Complexity from Partial Update

$$\min_{\boldsymbol{x}\in C\subset\mathbb{R}^d}\max_{\boldsymbol{y}\in\mathbb{R}^n}\{L(\boldsymbol{x},\boldsymbol{y})=f(\boldsymbol{x})+y^\top A\boldsymbol{x}-g(\boldsymbol{y})\}$$

(Ours) Maintain $\boldsymbol{w}=A\boldsymbol{x}$ and $\boldsymbol{z}=A^\top\boldsymbol{y}$.

For each iteration,

Operation	Cost
1. Compute full gradient $\partial_{\boldsymbol{x}}L=A^\top\boldsymbol{y}+f'(\boldsymbol{x})$ (Ours) $\Rightarrow\partial_{\boldsymbol{x}}L=\boldsymbol{z}+f'(\boldsymbol{x})$	$\mathcal{O}(nd)$ $\Rightarrow\mathcal{O}(d)$
2. Conduct Block FW (Eqn. (1)) on $\boldsymbol{x}$ to find s-sparse update $\Delta\boldsymbol{x}$	$\mathcal{O}(d)$
3. Update $\boldsymbol{x}^+\leftarrow(1-\eta)\boldsymbol{x}+\eta\Delta\boldsymbol{x}$ 3^+ (Ours) $\boldsymbol{w}^+\leftarrow(1-\eta)\boldsymbol{w}+\eta A\Delta\boldsymbol{x}$	$\mathcal{O}(d)$ $\mathcal{O}(sn)$
4. Greedy block-k coordinate ascent for $\boldsymbol{y}$ 4^+ (Ours) and maintain $\boldsymbol{y}$	$\mathcal{O}(nd)$ $\Rightarrow\mathcal{O}(kd)$

Remarks:
1. Take $k=ns/d$, the iteration complexity is $\mathcal{O}(sn)$.
2. The advantage comes from the fact that gradient could be maintained with the bilinear form.
3. The per iteration progress is as large as that of full projected gradient descent + gradient ascent steps

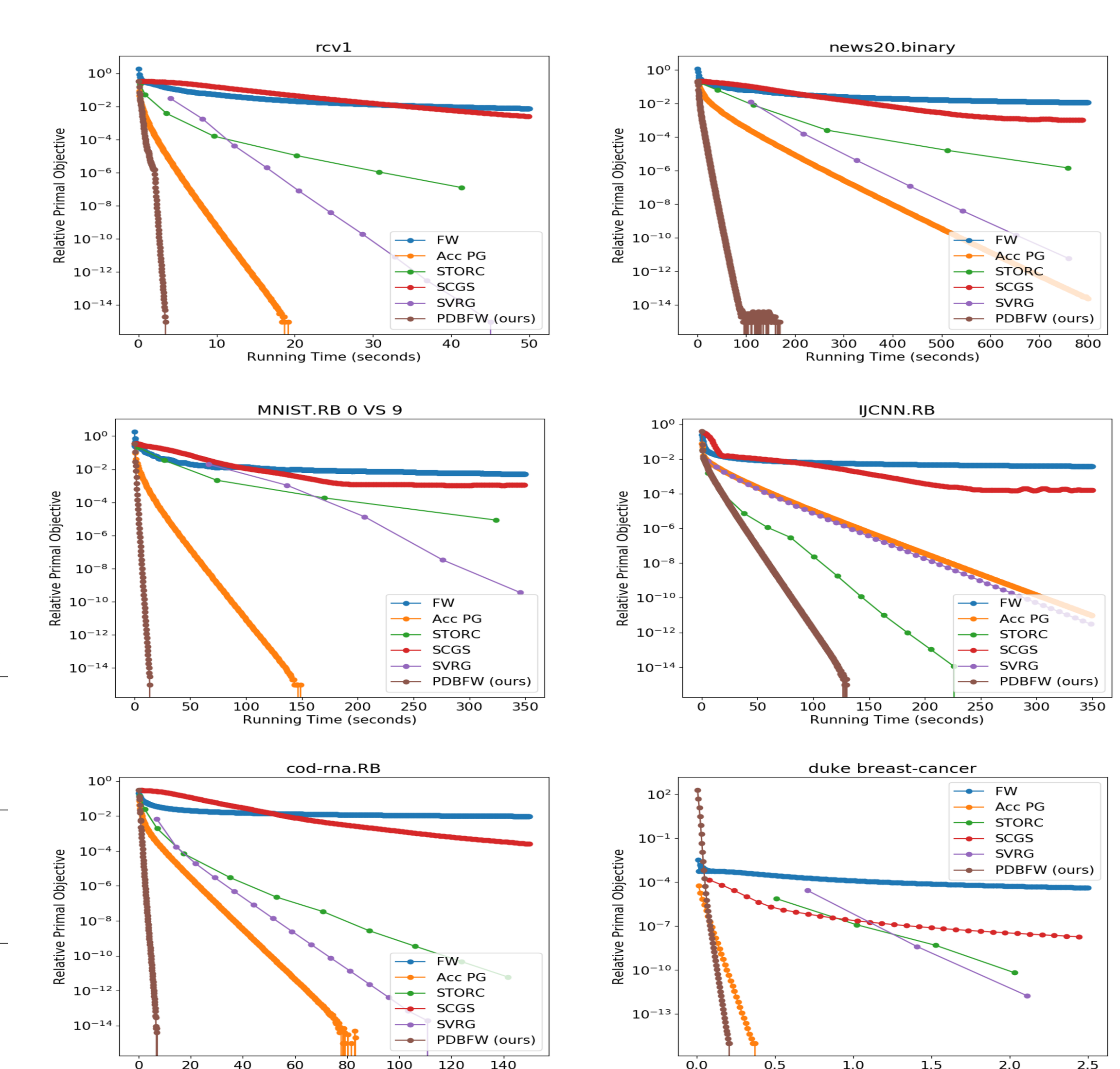
Time Comparisons

Algorithm	Per Iteration Cost	Iteration Complexity
Frank Wolfe	$\mathcal{O}(nd)$	$\mathcal{O}(\frac{1}{\epsilon})$
AGD	$\mathcal{O}(nd)$	$\mathcal{O}(\sqrt{\kappa}\log\frac{1}{\epsilon})$
SVRG	$\mathcal{O}(nd)$	$\mathcal{O}((1+\kappa/n)\log\frac{1}{\epsilon})$
SCGS	$\mathcal{O}(\kappa^2\frac{\#\text{iter}^3}{\epsilon^2}d)$	$\mathcal{O}(\frac{1}{\epsilon})$
STORC	$\mathcal{O}(\kappa^2 d+nd)$	$\mathcal{O}(\log\frac{1}{\epsilon})$
Ours	$\mathcal{O}(ns)$	$\mathcal{O}((1+\kappa/n)\log\frac{1}{\epsilon})$

Remarks:
1. s is the sparsity of primal optimal induced by ℓ_1 constraint.
2. For algorithm and complexity for nuclear norm constraints, refer to our paper to details.

Experimental Results

Compared methods:
(1) Accelerated ProjectedGradient Descent (Acc PG)
(2) Frank-Wolfe algorithm (FW)
(3) Stochastic Variance ReducedGradient (SVRG)
(4) Stochastic Conditional Gradient Sliding (SCGS)
(5) StochasticVariance-Reduced Conditional Gradient Sliding (STORC)



References

[1] 1. Qi Lei, et al. "Doubly greedy primal-dual coordinate descent for sparse empirical risk minimization." ICML-Volume 70. JMLR. org, 2017

[2] 2. Zeyuan Allen-Zhu, et al. "Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls." In Advances in Neural Information Processing Systems, 2017.

Our Algorithm: Primal-Dual Block Generalized Frank-Wolfe