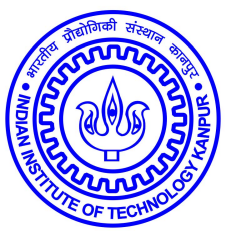
Projection Free Dynamic Online Learning

D. S. Kalhan*, A. S. Bedi†, A. Koppel†, K. Rajawat*, A. Gupta*, and A. Banerjee*

ICASSP 2020
*Indian Institute of Technology, Kanpur, India
† US Army Research Laboratory, Adelphi, MD, USA

April 11, 2020



Outline

Pivtol Points

Focused on developing Online Optimization Schemes

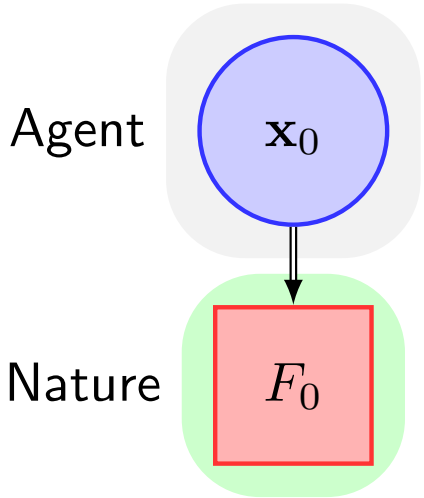
- Obviate the need of projection
  $\Rightarrow$ projection requires a quadratic problem to be solved at each step

- Robust to gradient estimation error
  $\Rightarrow$ exact online gradients may not be available
  $\Rightarrow$ dependence on unknown distributions or latency required for sampling

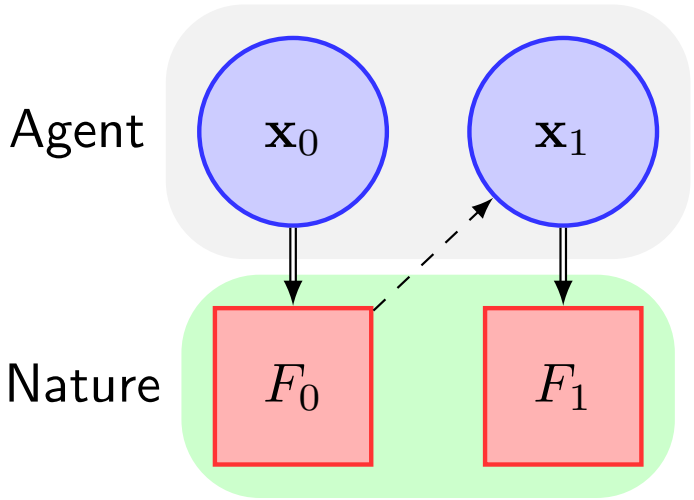
- Suitable for large scale networks
  $\Rightarrow$ exact gradient computations is expensive for large datasets[1].

[1] Lin Chen et al. "Projection-free online optimization with stochastic gradient: From convexity to submodularity". In: *arXiv preprint arXiv:1802.08183* (2018).

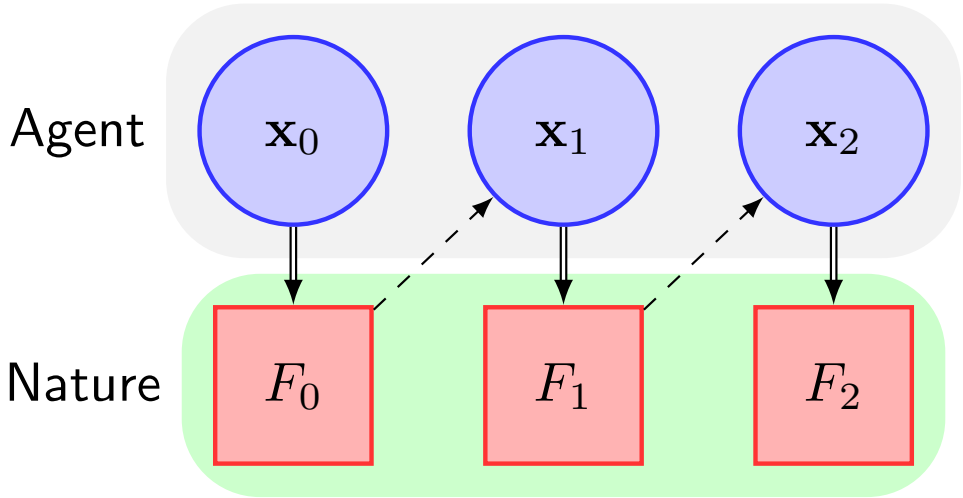
Online Learning



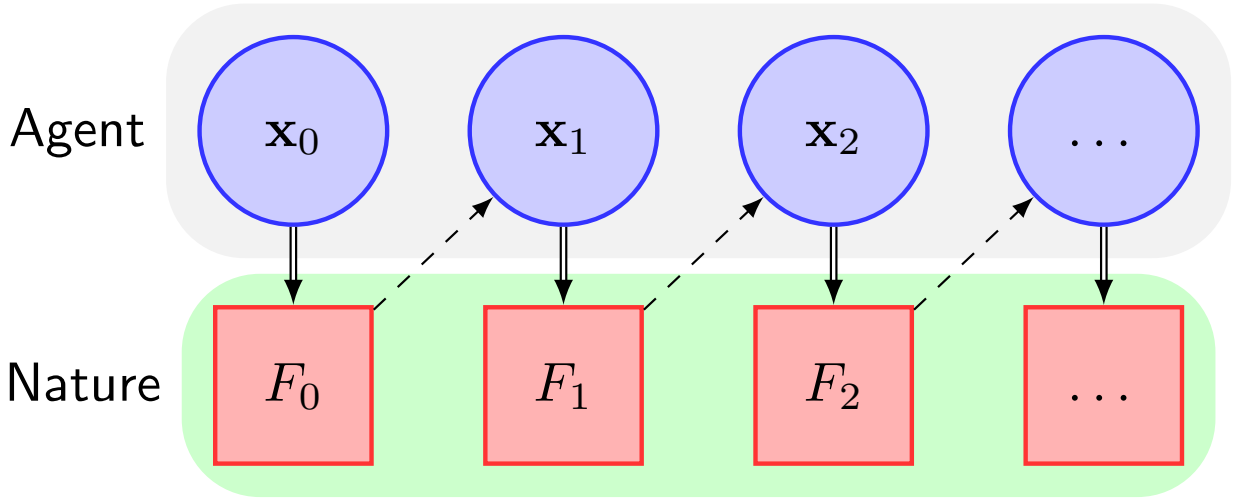
Online Learning



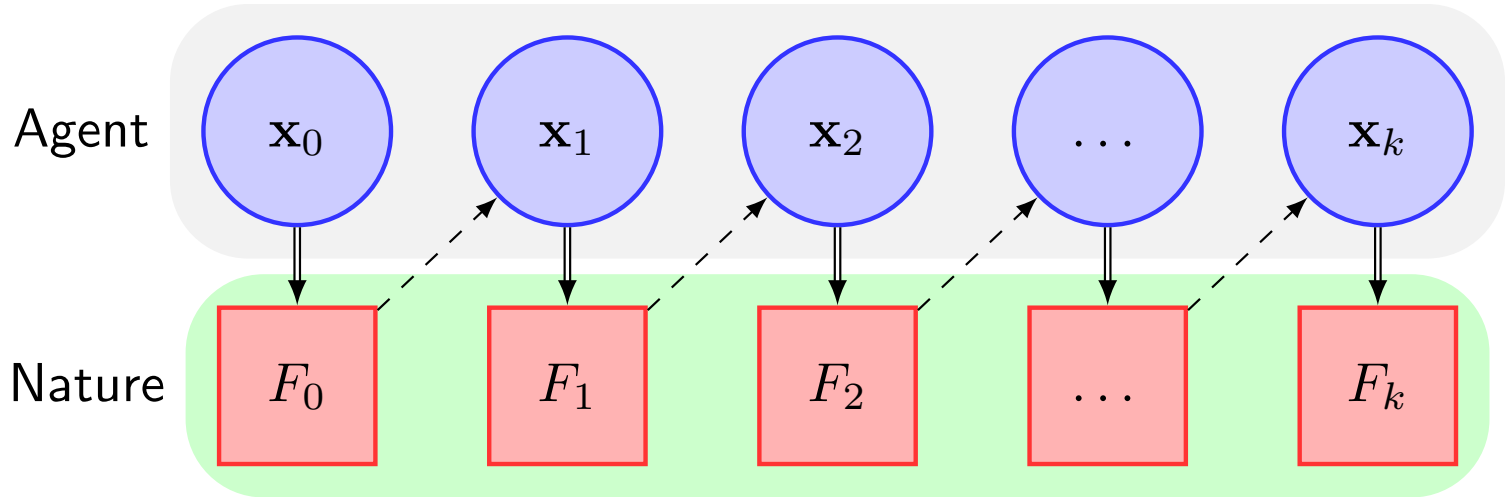
Online Learning



Online Learning



Online Learning



- Repeated game over a convex $\mathcal{X} \subset \mathbb{R}^n$
- At the t_{th} round, agent plays $\mathbf{x}_t \in \mathcal{X}$, Nature reveals $F_t : \mathcal{X} \to \mathbb{R}$
  $\Rightarrow$ Suffer arbitrary independent (antagonistic) convex loss $F_t(\mathbf{x}_t)$
- Static Regret $\Rightarrow$ performance metric for online learning

$$\mathbf{Reg}_T^S := \sum_{t=1}^{T} F_t(\mathbf{x}_t) - \sum_{t=1}^{T} F_t(\mathbf{x}^\star)$$

Dynamic Regret

- Dynamic Regret $\Rightarrow$ performance metric for online learning

$$\mathbf{Reg}_K := \sum_{t=1}^{T} F_t(\mathbf{x}_t) - \sum_{t=1}^{T} F_t(\mathbf{x}_t^\star)$$

- For fixed T, $\mathbf{x}_t^\star = \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x})$ is *offline* learner
  - $\Rightarrow$ Price for causal operation
  - $\Rightarrow$ How much we pay for not being clairvoyant

- Goal: $\mathbf{Reg}_T/T \to 0$ as $T \uparrow$, online gradient descent (Zinkevich, '03)
- Lower bounded by static regret

$$\min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} F_t(\mathbf{x}) \leq \sum_{t=1}^{T} F_t(\mathbf{x}_t^\star) \tag{1}$$

Dynamic Regret

Considered metrics of non-stationarity:

- Loss variation:

$$V_T = \sum_{t=1}^{T} \sup_{\mathbf{x} \in \mathcal{X}} |F_t(\mathbf{x}) - F_{t-1}(\mathbf{x})|.$$

- Gradient variation:

$$D_T = \sum_{t=1}^{T} \|\nabla F_t(\mathbf{x}_t) - \nabla F_{t-1}(\mathbf{x}_{t-1})\|^2$$

Our Approach

- Constraint satisfaction at each time slot
  $\Rightarrow$ Projection requires a quadratic problem to be solved at each step[2].

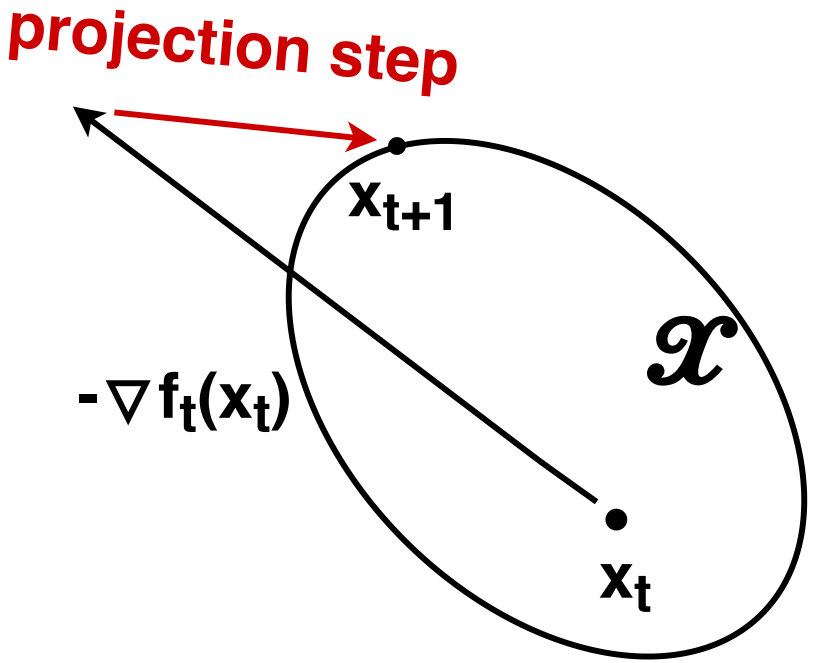


Figure 1: Projection Step in Online Gradient Descent Algorithm

[2] R Tyrrell Rockafellar. "Monotone operators and the proximal point algorithm". In: *SIAM journal on control and optimization* 14.5 (1976), pp. 877–898.

Our Approach

- Avoid projections in online constrained settings
  $\Rightarrow$ Frank-Wolfe (conditional gradient) method $\Rightarrow$ moving in feasible direction that is collinear with the gradient through the solution of a linear program[3].

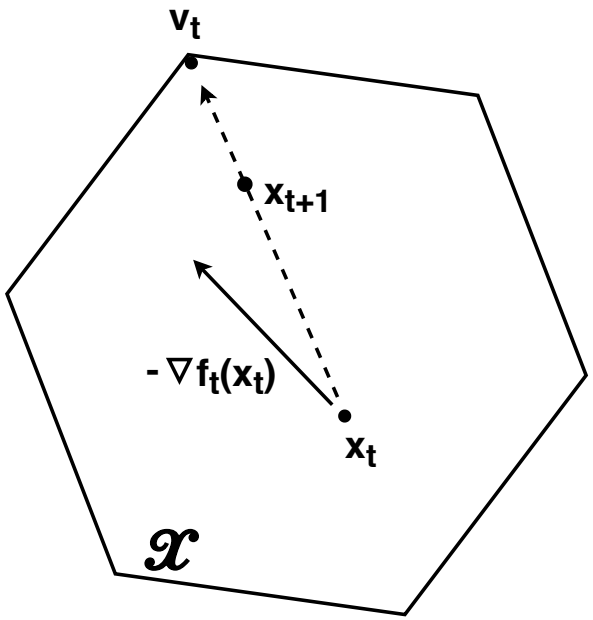


Figure 2: Frank Wolfe Algorithm

[3] Marguerite Frank and Philip Wolfe. "An algorithm for quadratic programming". In: *Naval research logistics quarterly* 3.1-2 (1956), pp. 95–110.

Related Work

Reference	Loss function.	Step-size	Batch	Regret definition	Rate
[Hazan at al][4]	$(L/D)t^{-1/4}$-strongly convex	diminishing	$\mathcal{O}(t)$	$\sum_{t=1}^{T} F(\mathbf{x}_t) - F(\mathbf{x}^\star)$	$\mathcal{O}\left(T^{3/4}\right)$
[Aryan et al][5]	convex	diminishing	$\mathcal{O}(1)$	$\mathbb{E}\left[F(\mathbf{x}_T) - F(\mathbf{x}^\star)\right]$	$\mathcal{O}(1/T^{1/3})$
[Shahim et al][6]	convex	depends on σ_2 & C_T	-	$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[F_t(\mathbf{x}_{i,t}) - F_t(\mathbf{x}_t^\star)\right]$	$\mathcal{O}\left(\sqrt{\frac{(1+C_T)T}{1-\sigma_2(W)}}\right)$
[Hazan et al][7]	1-strongly convex	diminishing	$\mathcal{O}(1)$	$\sum_{t=1}^{T}\left[F_t(\mathbf{x}_t) - F_t(\mathbf{x}^\star)\right]$	$\mathcal{O}\left(T^{3/4}\right)$
This work	convex	constant	$\mathcal{O}(1)$	$\sum_{t=1}^{T} F_t(\mathbf{x}_t) - F_t(\mathbf{x}_t^\star)$	$\mathcal{O}\left(\sqrt{T}\left(1 + V_T + \sqrt{D_T}\right)\right)$
This work	convex (partial feedback)	constant	$\mathcal{O}(1)$	$\sum_{t=1}^{T}\mathbb{E}\left[F_t(\mathbf{x}_t) - F_t(\mathbf{x}_t^\star)\right]$	$\mathcal{O}\left(1 + T^{\frac{5}{6}} + \sqrt{T}V_T + T^{\frac{5}{6}}\sqrt{D_T}\right)$

Table 1: Summary of the related works compared to the present work.

[4] Elad Hazan and Satyen Kale. "Projection-free Online Learning". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, 2012, pp. 1843–1850. ISBN: 978-1-4503-1285-1. URL: http://dl.acm.org/citation.cfm?id=3042573.3042808.

[5] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. "Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization". In: *arXiv preprint arXiv:1804.09554* (2018).

[6] Shahin Shahrampour and Ali Jadbabaie. "Distributed online optimization in dynamic environments using mirror descent". In: *IEEE Transactions on Automatic Control* 63.3 (2018), pp. 714–725.

[7] Elad Hazan et al. "Introduction to online convex optimization". In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325.

Contributions

- Online Frank Wolfe (OFW)
  - $\Rightarrow$ Projection-free
  - $\Rightarrow$ Exact gradient

- Online Frank Wolfe under partial feedback(OFW-inexact)
  - $\Rightarrow$ Projection-free
  - $\Rightarrow$ Noisy gradient

Online Frank Wolfe Algorithm

Online Frank Wolfe (OFW)

Algorithm 1 Online Frank-Wolfe Algorithm (OFW)

1: **Require** step size $0 < \gamma < 1$.
2: **Initialize** $t = 0$, $d_0 = 0$ and choose $x_0 \in \mathcal{X}$.
3: **for** t = 1,2.....**do**.
4: **Compute** the gradient $\nabla F_t(\mathbf{x}_t)$.
5: **Compute** $\mathbf{v}_t = \arg\min_{\mathbf{v} \in \mathcal{X}} \langle \nabla F_t(\mathbf{x}_t), \mathbf{v} \rangle$
6: **Update** $\mathbf{x}_{t+1} = (1 - \gamma)\hat{\mathbf{x}}_t + \gamma \mathbf{v}_t$.
7: **end for**

Technical Assumptions

- **Assumption 1** : The set $\mathcal{X}$ is convex and compact with diameter D, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, it holds that $\|\mathbf{x} - \mathbf{y}\| \leq D$.

- **Assumption 2** : The gradient of loss $\nabla F_t(\cdot)$ is Lipschitz with parameter L_1, which implies that

$$\|\nabla F_t(\mathbf{x}) - \nabla F_t(\mathbf{y})\| \leq L_1 \|\mathbf{x} - \mathbf{y}\| \text{ for all } t \text{ and } (\mathbf{x}, \mathbf{y}) \in \mathcal{X}. \quad (2)$$

Technical Assumptions

- **Assumption 1** : The set $\mathcal{X}$ is convex and compact with diameter D, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, it holds that $\|\mathbf{x} - \mathbf{y}\| \leq D$.

- **Assumption 2** : The gradient of loss $\nabla F_t(\cdot)$ is Lipschitz with parameter L_1, which implies that

$$\|\nabla F_t(\mathbf{x}) - \nabla F_t(\mathbf{y})\| \leq L_1 \|\mathbf{x} - \mathbf{y}\| \text{ for all } t \text{ and } (\mathbf{x}, \mathbf{y}) \in \mathcal{X}. \quad (2)$$

Regret Bounds

Theorem

When loss functions F_t are **convex**, *under step-size selection $\gamma = \frac{1}{\sqrt{T}}$, we have*

$$Reg_T^D \leq \mathcal{O}\left(\sqrt{T}\left(1 + V_T + \sqrt{D_T}\right)\right), \tag{3}$$

where,

$$D_T = \sum_{t=1}^{T} \left\|\nabla F_t(\mathbf{x}_t) - \nabla F_{t-1}(\mathbf{x}_{t-1})\right\|^2 , \quad V_T = \sum_{t=1}^{T} \sup_{\mathbf{x}\in\mathcal{X}} |F_t(\mathbf{x}) - F_{t-1}(\mathbf{x})|.$$

Online Frank Wolfe under partial feedback

Our Approach

- Estimate gradient from partial/noisy gradient.

$\Rightarrow$ reduce error in gradient approximation using gradient averaging technique[8] defined as

$$\mathbf{d}_t = (1-\rho)\mathbf{d}_{t-1} + \rho\nabla f_t(\mathbf{x}_t, \mathbf{z}_t). \tag{4}$$

where, $\mathbf{d}_t$ is the estimated gradient
$\nabla f_t(\mathbf{x}_t, \mathbf{z}_t)$ is partial/noisy gradient
ρ is inertia factor.

[8] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. "Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization". In: *arXiv preprint arXiv:1804.09554* (2018).

Online Frank-Wolfe under partial feedback

Algorithm 2 Online Frank-Wolfe under partial feedback (OFW-inexact)

1: **Require** step sizes $0 < \rho < 1$ and $0 < \gamma < 1$.
2: **Initialize** $t = 0$, $d_0 = 0$ and choose $\mathbf{x}_0 \in \mathcal{X}$.
3: **for** t = 1,2.....**do**
4: **Update** the gradient estimate $\mathbf{d}_t = (1-\rho)\mathbf{d}_{t-1} + \rho\nabla f_t(\mathbf{x}_t, \mathbf{z}_t)$.
5: **Compute** $\mathbf{v}_t = \arg\min_{\mathbf{v}\in\mathcal{X}} \langle \mathbf{d}_t, \mathbf{v} \rangle$.
6: **Update** $\mathbf{x}_{t+1} = (1-\gamma)\mathbf{x}_t + \gamma\mathbf{v}_t$.
7: **end for**

Technical Assumptions

- **Assumption 3:** The variance of the unbiased stochastic gradients $\nabla f_t(\mathbf{x}, \mathbf{z})$ is bounded by σ^2.

$$\mathbb{E}\big[\, \|\nabla f_t(\mathbf{x}, \mathbf{z}) - \nabla F_t(\mathbf{x})\|^2 \,\big] \leq \sigma^2, \qquad \text{for all } \; t. \tag{5}$$

Regret Bounds

Theorem

When F_t is convex, under step-size and inertia selections $\gamma = \frac{1}{\sqrt{T}}$, $\rho = \frac{1}{\sqrt[3]{T}}$, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[F_t(\mathbf{x}_t)\right] - \sum_{t=1}^{T} F_t(\mathbf{x}_t^\star) \leq \quad \mathcal{O}\left(1 + T^{(5/6)} + \sqrt{T}V_T + +T^{(5/6)}\sqrt{D_T}\right)$$

where,

$$D_T = \sum_{t=1}^{T} \left\|\nabla F_t(\mathbf{x}_t) - \nabla F_{t-1}(\mathbf{x}_{t-1})\right\|^2 \ , \quad V_T = \sum_{t=1}^{T} \sup_{\mathbf{x}\in\mathcal{X}} |F_t(\mathbf{x}) - F_{t-1}(\mathbf{x})|.$$

Experimental Results

Online Matrix Completion

- Best possible low rank approximation $\mathbf{X} \in \mathbb{R}^{m \times n}$ of a given matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$.

  $\Rightarrow$ Observed entries of fixed size arrive from random locations of $\mathbf{M}$ as $OB_1, OB_2, ..., OB_T$.

  $\Rightarrow$ The problem is defined as[9]

$$\min_{\mathbf{X}_{ij}} \sum_{(ij) \in OB} (\mathbf{X}_{ij} - \mathbf{M}_{ij})^2 \quad \text{such that } \|\mathbf{X}\|_* \leq k. \tag{6}$$

  we have considered only 25% of the samples full gradient from random locations.

[9] Elad Hazan et al. "Introduction to online convex optimization". In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325, Chap. 7.

Online Matrix Completion

- Best possible low rank approximation $\mathbf{X} \in \mathbb{R}^{m \times n}$ of a given matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$.
  $\Rightarrow$ Observed entries of fixed size arrive from random locations of $\mathbf{M}$ as $OB_1, OB_2, ..., OB_T$.
  $\Rightarrow$ The problem is defined as[9]

$$\min_{\mathbf{X}_{ij}} \sum_{(ij) \in OB} (\mathbf{X}_{ij} - \mathbf{M}_{ij})^2 \quad \text{such that } \|\mathbf{X}\|_* \leq k. \tag{6}$$

  we have considered only 25% of the samples full gradient from random locations.

[9] Elad Hazan et al. "Introduction to online convex optimization". In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325, Chap. 7.

Online Matrix Completion

- Best possible low rank approximation $\mathbf{X} \in \mathbb{R}^{m \times n}$ of a given matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$.
  $\Rightarrow$ Observed entries of fixed size arrive from random locations of $\mathbf{M}$ as $OB_1, OB_2, ..., OB_T$.
  $\Rightarrow$ The problem is defined as[9]

$$\min_{\mathbf{X}_{ij}} \sum_{(ij) \in OB} (\mathbf{X}_{ij} - \mathbf{M}_{ij})^2 \quad \text{such that } \|\mathbf{X}\|_* \leq k. \tag{6}$$

  we have considered only 25% of the samples full gradient from random locations.

[9] Elad Hazan et al. "Introduction to online convex optimization". In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325, Chap. 7.

Online Matrix Completion

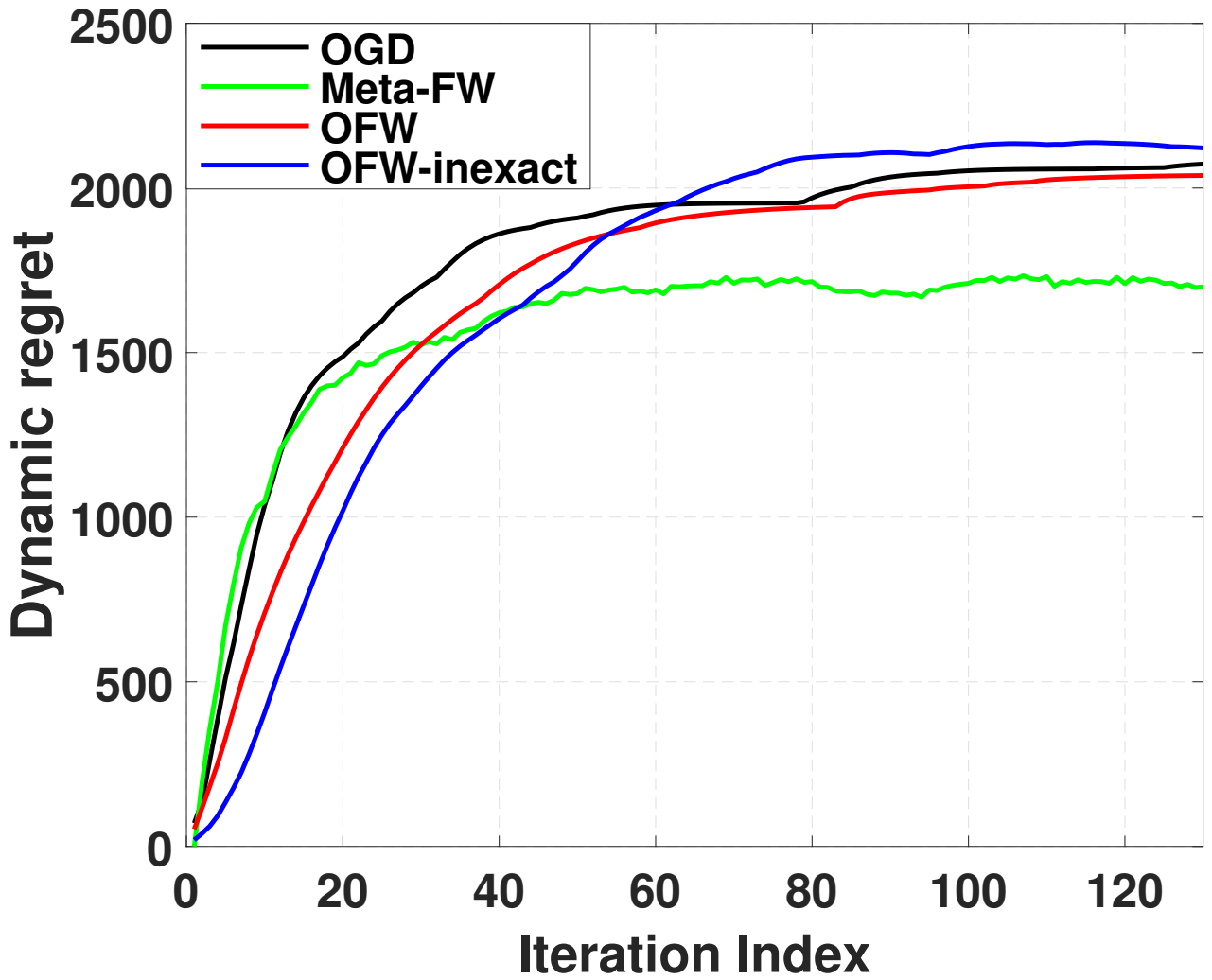


Figure 3: Dynamic regret on online matrix completion

Online Matrix Completion

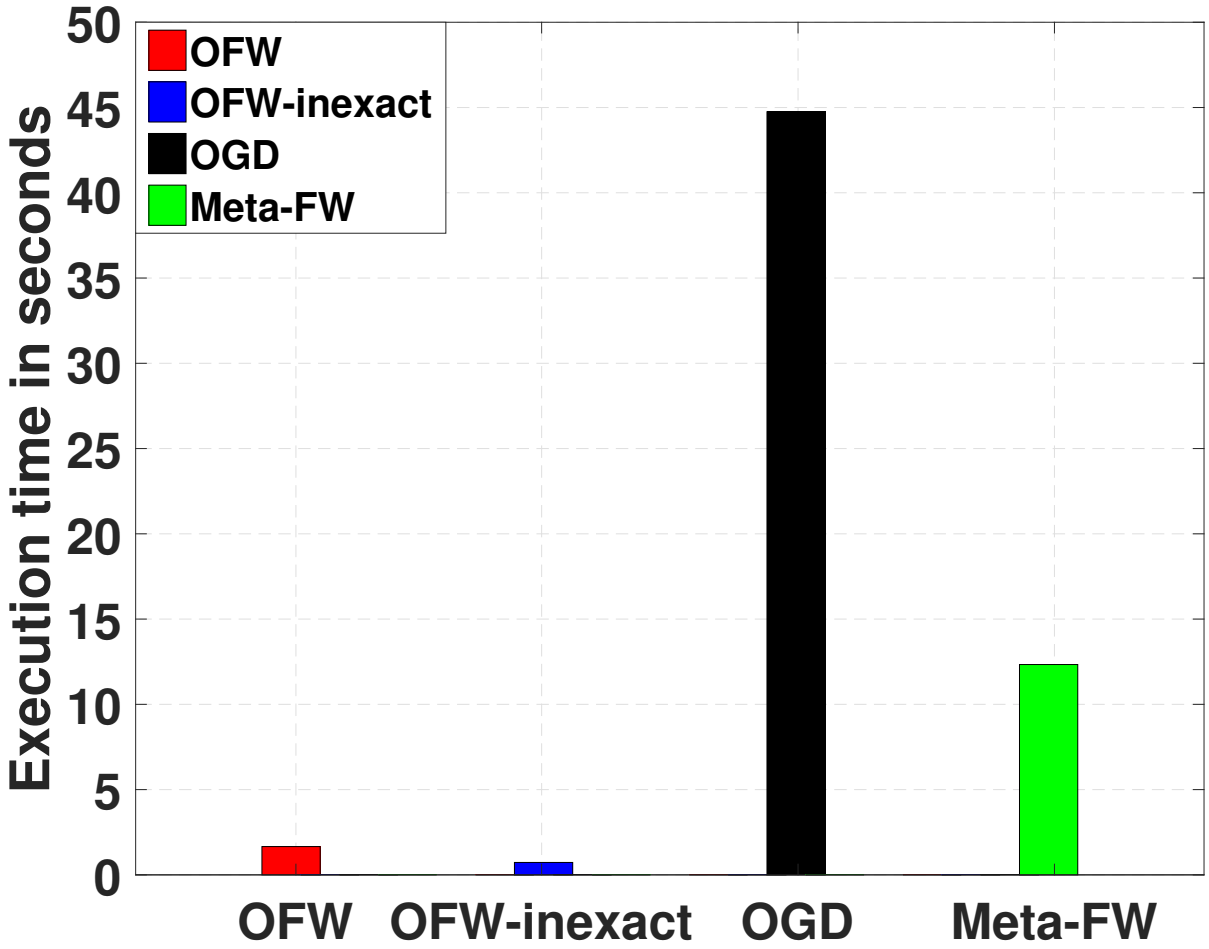


Figure 4: Execution time comparison on online matrix completion

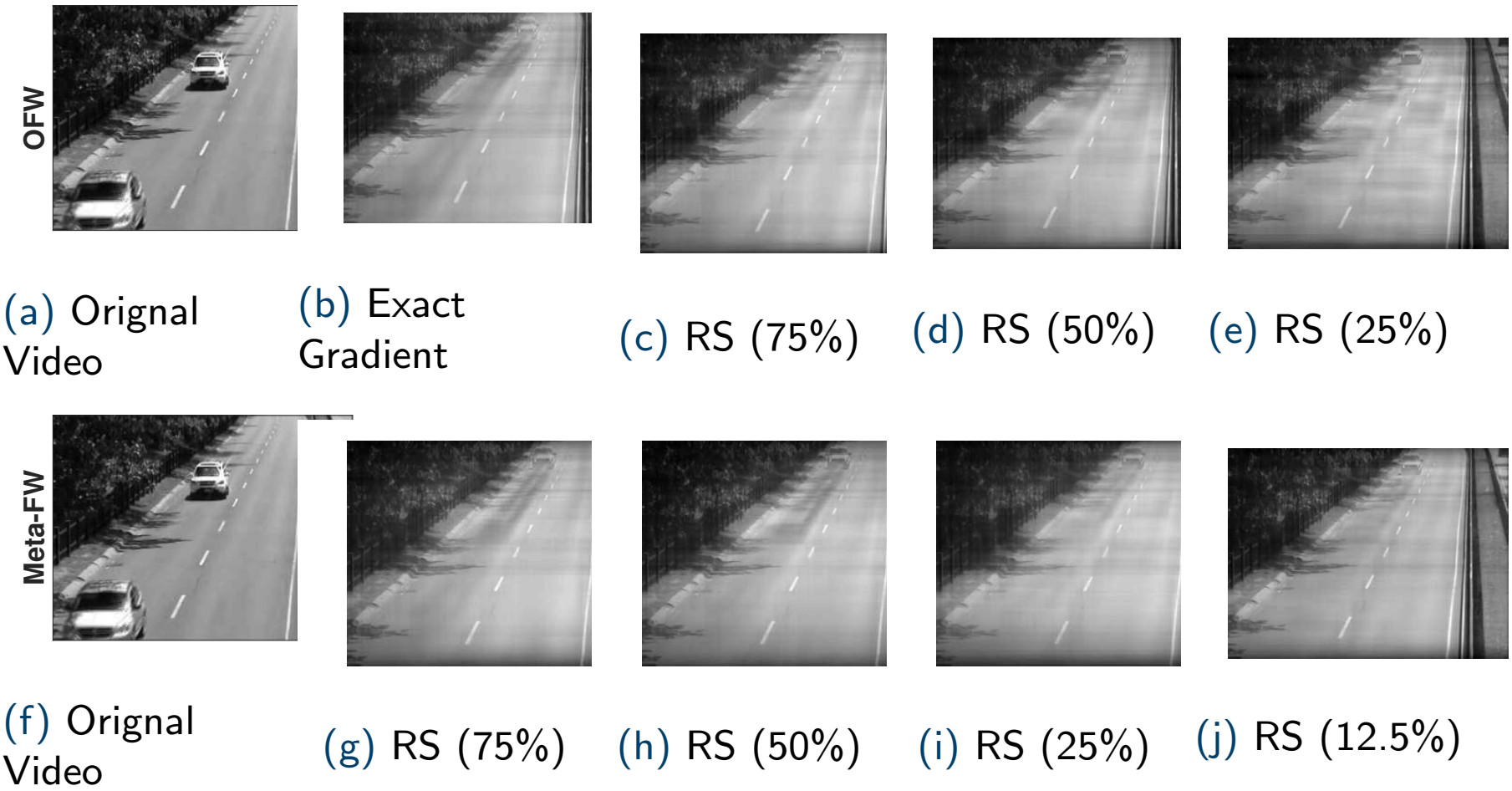
Background extraction problem

- Extension of matrix completion problem on real dataset[10]
  $\Rightarrow$ At each instant we observe a video frame and collect it into matrix $\mathbf{M}_t$.
  $\Rightarrow$ The goal of the problem is to extract the background from the video which is conceptually the low-rank estimate $\mathbf{M}_t$ of the data matrix.
  $\Rightarrow$ The problem can be formulated as:

$$\min_{\mathbf{L}_t} \|\mathbf{M}_t - \mathbf{L}_t\|_F^2 + \frac{1}{2}\|\mathbf{L}_t\|_F^2$$
$$\text{such that } \|\mathbf{L}_t\|_* \leq k$$

[10] Nil Goyette et al. "Changedetection. net: A new change detection benchmark dataset". In: *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE. 2012, pp. 1–8.

Background extraction problem



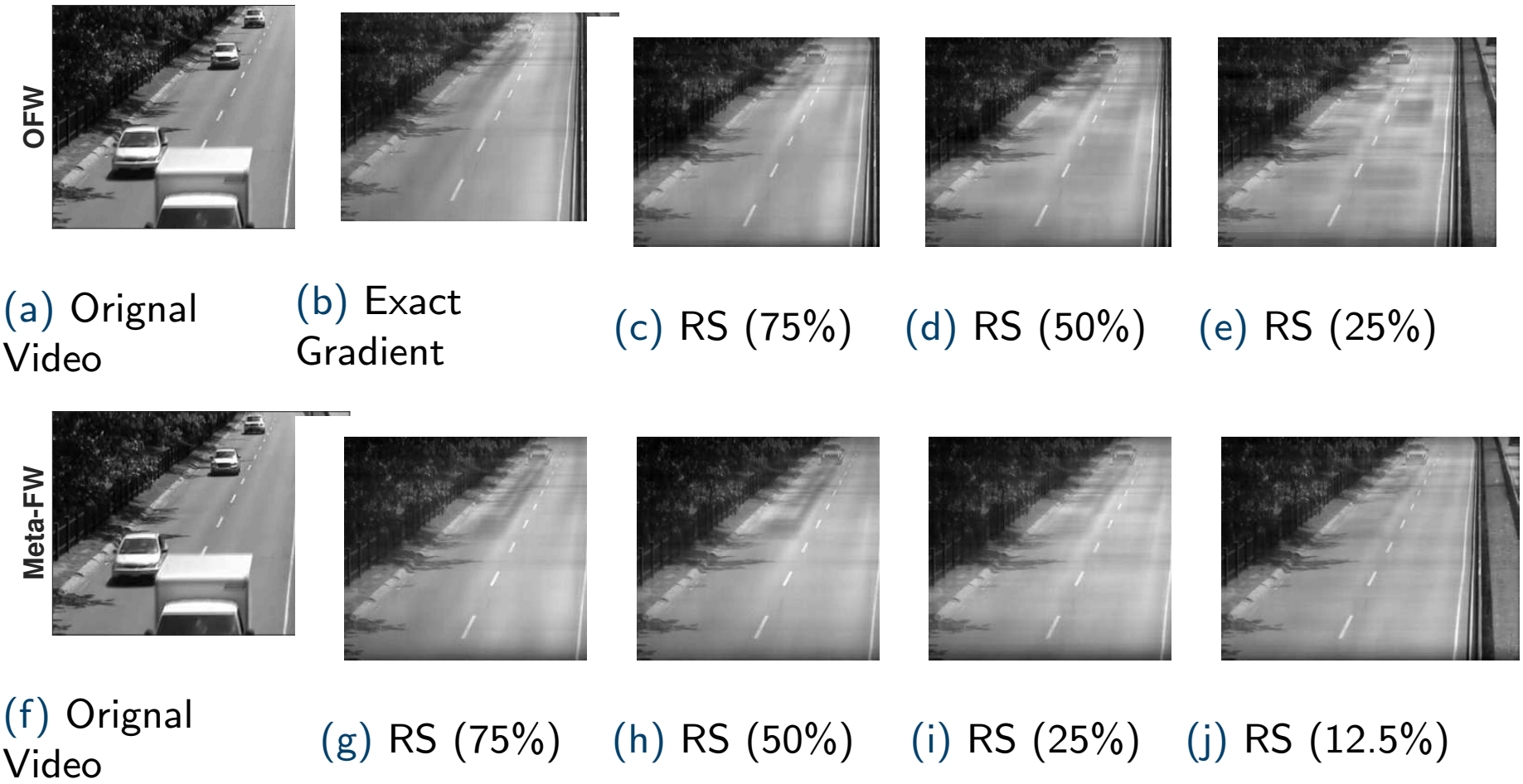
(a) Orignal Video (b) Exact Gradient (c) RS (75%) (d) RS (50%) (e) RS (25%)

(f) Orignal Video (g) RS (75%) (h) RS (50%) (i) RS (25%) (j) RS (12.5%)

Figure 5: Background Extraction Problem on Highway data set at 1st instant

RS denotes random sampling.

Background extraction problem



(a) Orignal Video (b) Exact Gradient (c) RS (75%) (d) RS (50%) (e) RS (25%)

(f) Orignal Video (g) RS (75%) (h) RS (50%) (i) RS (25%) (j) RS (12.5%)

Figure 6: Background Extraction Problem on Highway data set at 2nd instant

RS denotes random sampling.

Background extraction problem

Algorithm	Exact Gradient	RS(75%)	RS(50%)	RS(25%)
OFW	4.6436	4.2325	3.1949	3.1396
Meta-FW	26.5808	26.5810	22.8794	21.1206

Table 2: Summary of execution time in seconds.

Experimental Results

(Background Extraction)

Online Frank Wolfe algorithm

Thank You