First order methods

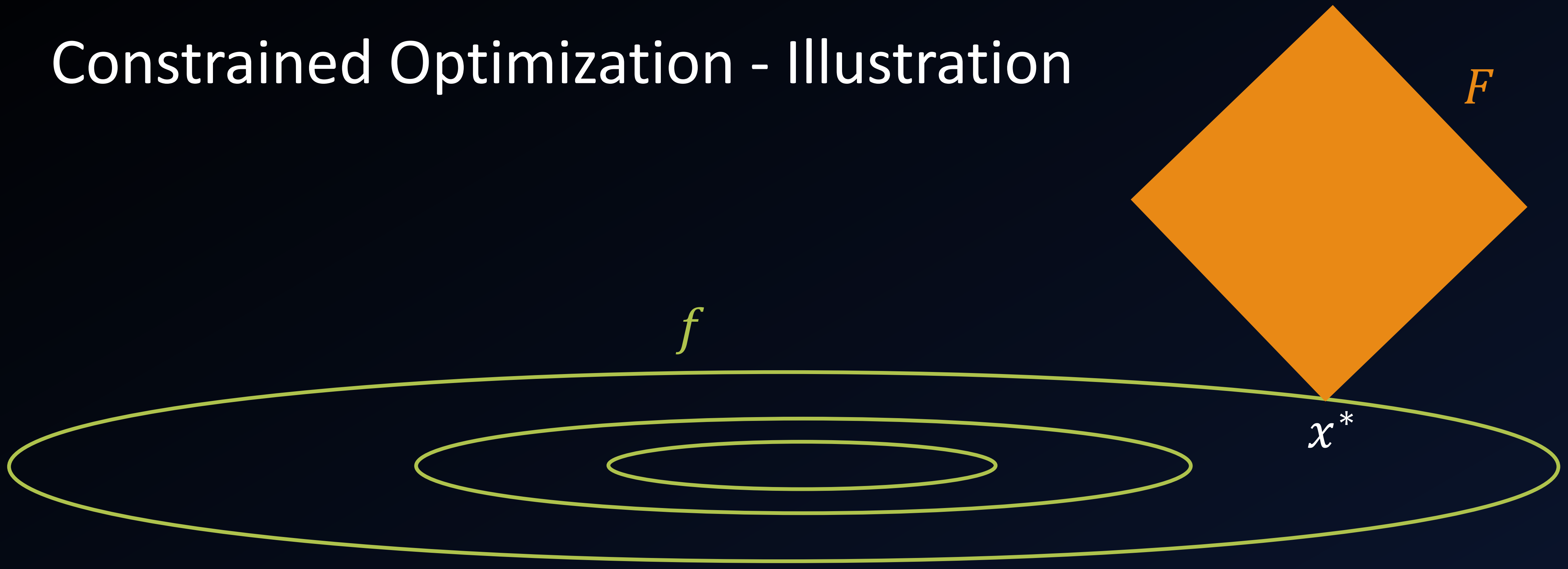
FOR CONVEX OPTIMIZATION

Saketh (IIT Bombay)

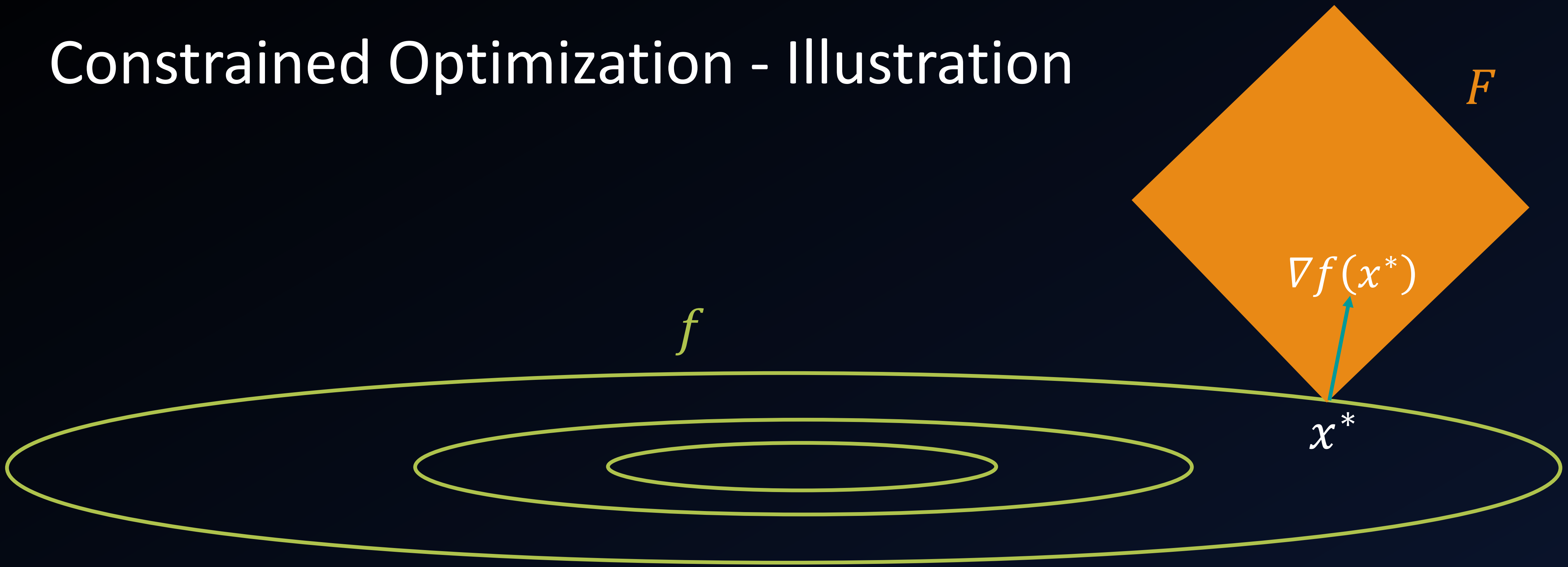
Topics

- Part – I
  - *Optimal* methods for unconstrained convex programs
    - Smooth objective
    - Non-smooth objective
- Part – II
  - *Optimal* methods for constrained convex programs
    - Projection based
    - Frank-Wolfe based
    - Functional constraint based
  - Prox-based methods for structured non-smooth programs

Constrained Optimization - Illustration



Constrained Optimization - Illustration



$$x^* \text{ is optimal} \Leftrightarrow \nabla f(x^*)^T u \geq 0 \;\forall\, u \in T_F(x^*)$$

Two Strategies

- Stay feasible and minimize
    - Projection based
    - Frank-Wolfe based

Two Strategies

- Alternate between
  - Minimization
  - Move towards feasibility set

Projection Based Methods

CONSTRAINED CONVEX PROGRAMS

Projected Gradient Method

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

X is closed convex

- $x_{k+1} = \text{argmin}_{x\in X} f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2s_k}\|x - x_k\|^2$

  $= \text{argmin}_{x\in X} \|x - (x_k - s_k\nabla f(x_k))\|^2$

  $= \Pi_X(x_k - s_k\nabla f(x_k))$

Projected Gradient Method

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

- $x_{k+1} = \text{argmin}_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2$

  $= \text{argmin}_{x \in X} \|x - (x_k - s_k \nabla f(x_k))\|^2$

  $= \Pi_X(x_k - s_k \nabla f(x_k))$

Projected Gradient Method

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

- $x_{k+1} = \mathrm{argmin}_{x\in X}\, f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2s_k}\|x - x_k\|^2$

  $= \mathrm{argmin}_{x\in X}\, \|x - (x_k - s_k \nabla f(x_k))\|^2$

  $\equiv \Pi_X(x_k - s_k \nabla f(x_k))$

X is simple:
oracle for projections

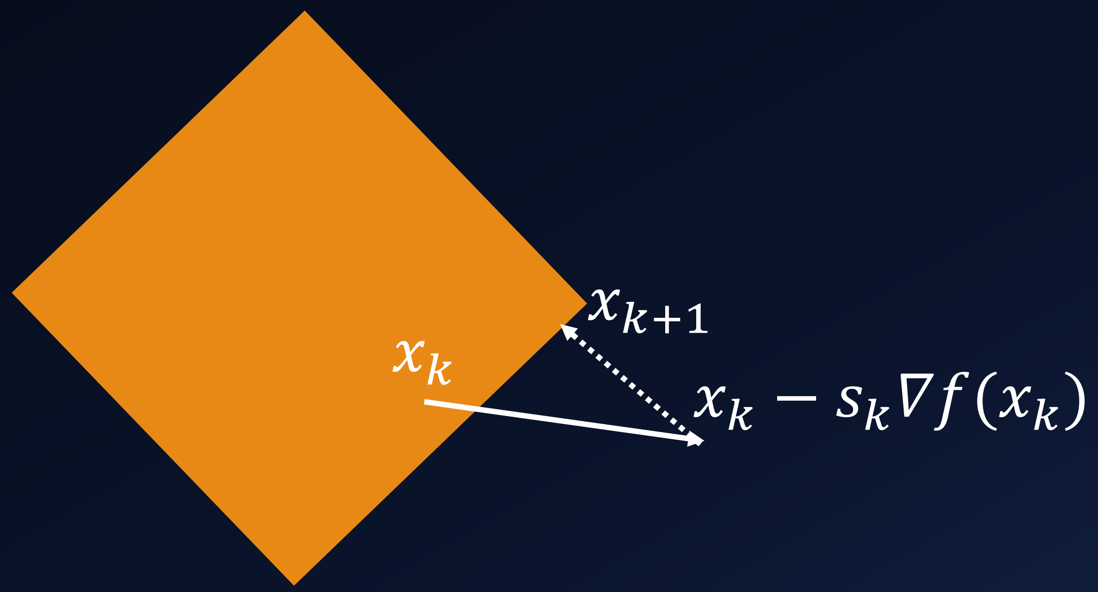
Projected Gradient Method

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

- $x_{k+1} = \operatorname{argmin}_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2$

$$= \operatorname{argmin}_{x \in X} \|x - (x_k - s_k \nabla f(x_k))\|^2$$

$$\equiv \Pi_X(x_k - s_k \nabla f(x_k))$$



Will it work?

- $$\|x_{k+1} - x^*\|^2 = \|\Pi_X(x_k - s_k\nabla f(x_k)) - x^*\|^2$$
  $$\leq \|(x_k - s_k\nabla f(x_k)) - x^*\|^2$$ (Why?)
- Remaining analysis exactly same (smooth/non-smooth)

- Analysis a bit more involved for projected accelerated gradient
  - Define gradient map: $h(x_k) = \frac{x_k - \Pi_X(x_k - s_k\nabla f(x_k))}{s_k}$
  - Satisfies same fundamental properties as gradient!

Will it work?

- $\|x_{k+1} - x^*\|^2 = \|\Pi_X(x_k - s_k \nabla f(x_k)) - x^*\|^2$

  $\leq \| (x_k - s_k \nabla f(x_k)) - x^*\|^2$ (Why?)

- Remaining analysis exactly same (smooth/non-smooth)

- Analysis a bit more involved for projected accelerated gradient
  - Define gradient map: $h(x_k) \equiv \frac{x_k - \Pi_X(x_k - s_k \nabla f(x_k))}{s_k}$
  - Satisfies same fundamental properties as gradient!

Simple sets

- Non-negative orthant
- Ball, ellipse
- Box, simplex
- Cones
- PSD matrices
- Spectrahedron

Summary of Projection Based Methods

- Rates of convergence remain exactly same
- Projection oracle needed (simple sets)
  - Caution with non-analytic cases

Frank-Wolfe Methods

CONSTRAINED CONVEX PROGRAMS

Avoid Projections

- $y_{k+1} = \text{argmin}_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \cancel{\frac{1}{2s_k} \|x - x_k\|^2}$

  $= \text{argmin}_{x \in X} \nabla f(x_k)^T x$ (Support Function)

- Restrict moving far away:
  - $x_{k+1} = s_k y_{k+1} + (1 - s_k) x_k$

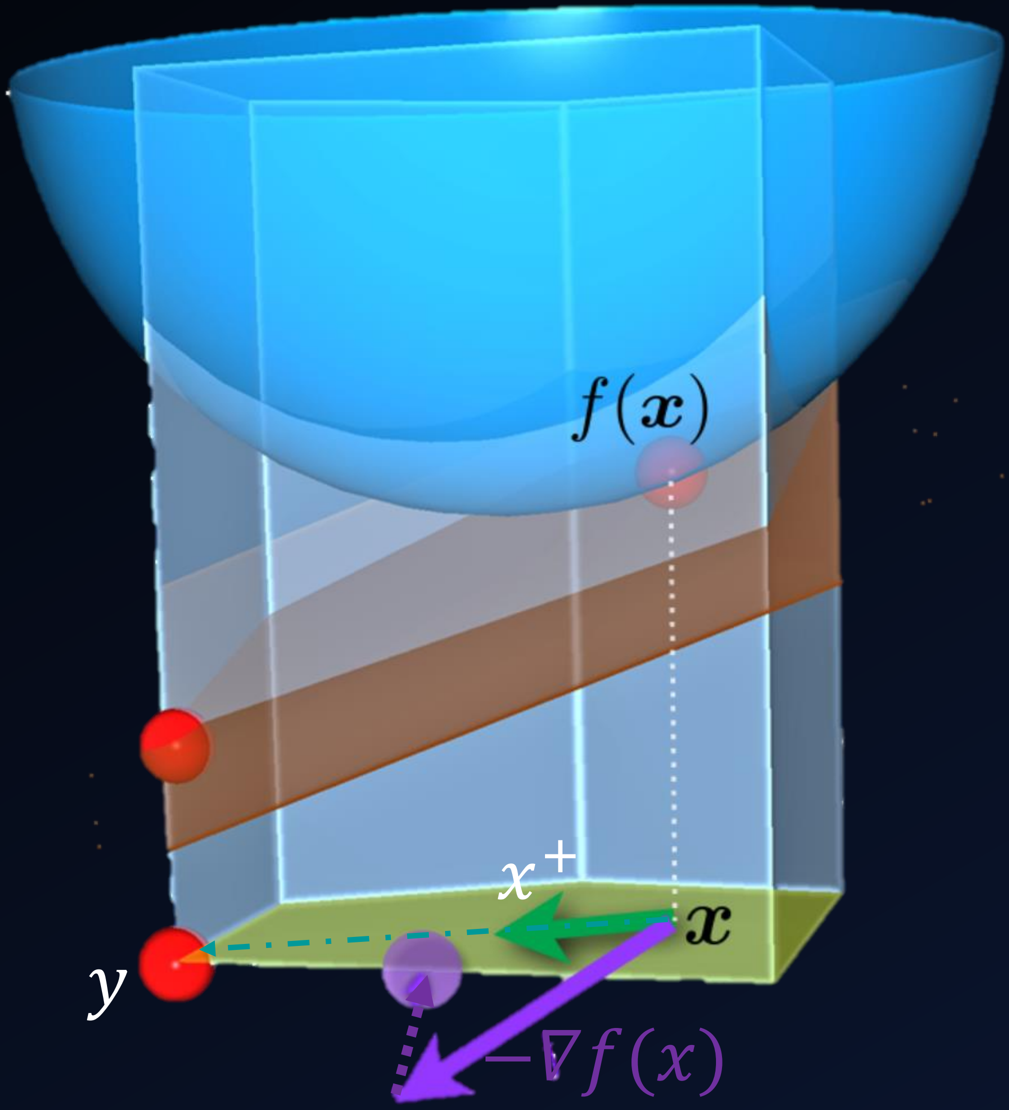
Avoid Projections [FW59]

- $y_{k+1} = \text{argmin}_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \cancel{\frac{1}{2s_k} \|x - x_k\|^2}$

  $= \text{argmin}_{x \in X} \nabla f(x_k)^T x$ (Support Function)

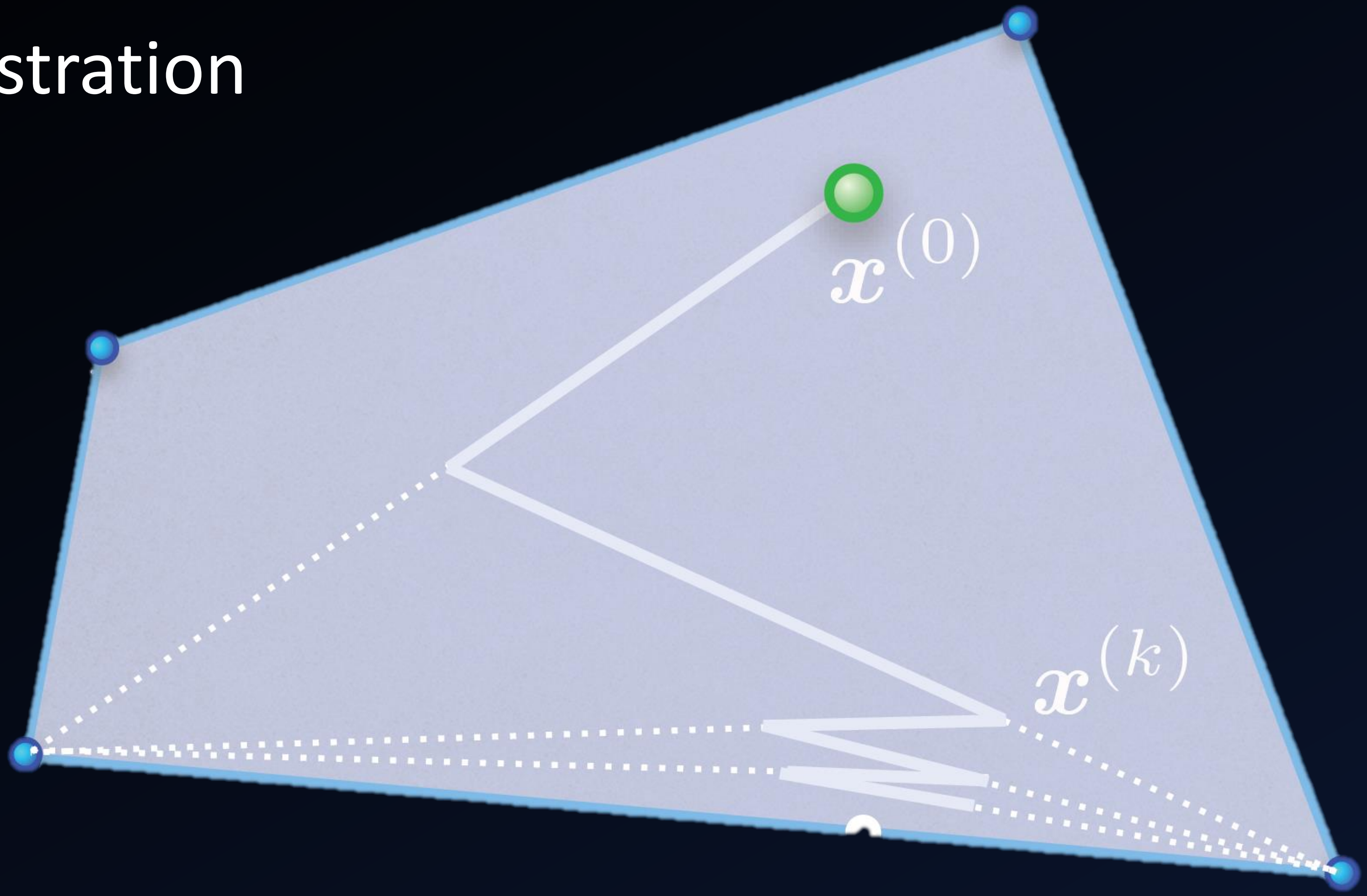
- Restrict moving far away:
  - $x_{k+1} \equiv s_k y_{k+1} + (1 - s_k) x_k$

Illustration



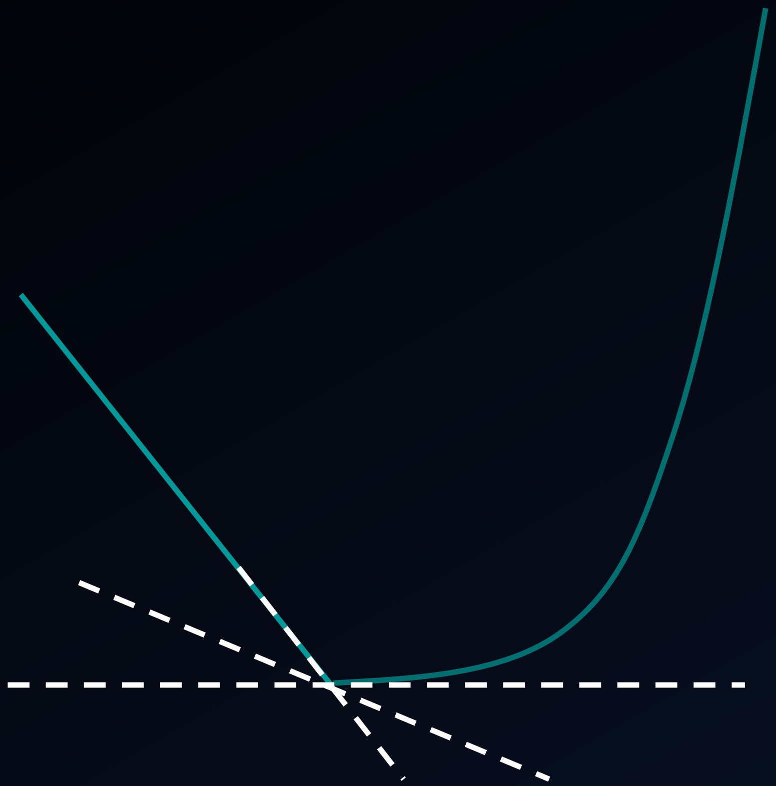
[Mart Jaggi, ICML 2014]

Illustration



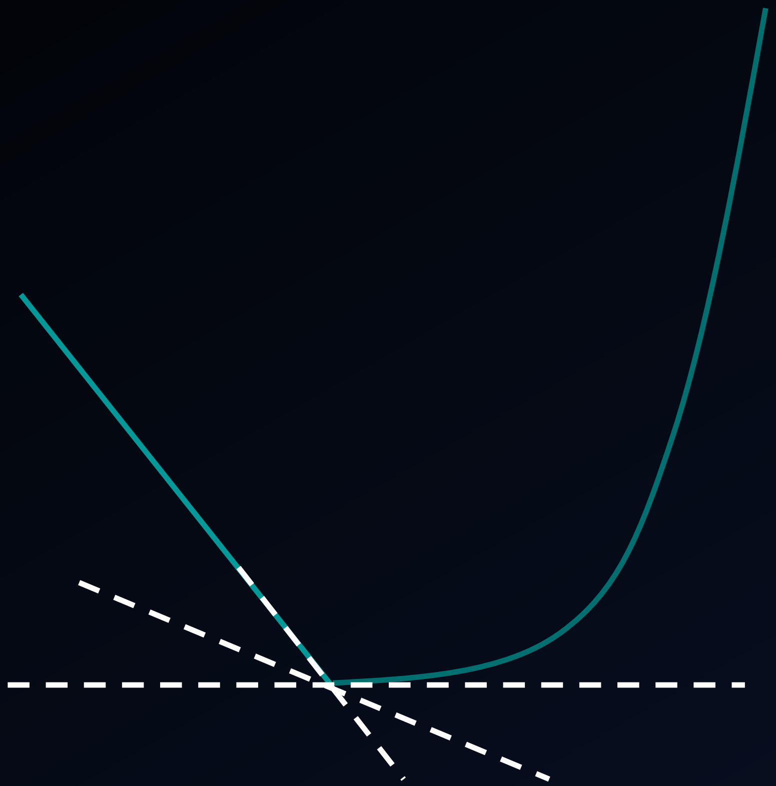
[Mart Jaggi, ICML 2014]

On Conjugates and Support Functions



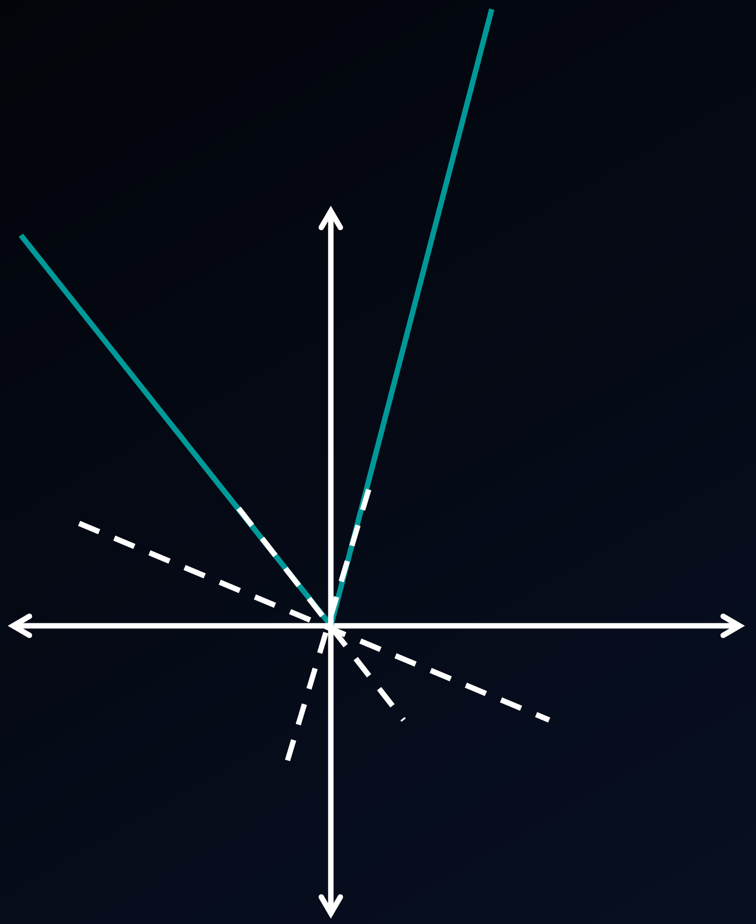
- Convex f is point-wise maximum of affine minorants
- Provides dual definition:
  - $f(x) = \max_{y \in Y} a_y^T x - b_y$, equivalently:
  - $\exists f^* \ni f(x) = \max_{y \in dom\, f^*} y^T x - f^*(y)$
  - f^* is called conjugate or Fenchel dual
- If $f^*(y)$ is indicator of set S we get conic f:
  - $f(x) = \max_{y \in S} y^T x$

On Conjugates and Support Functions



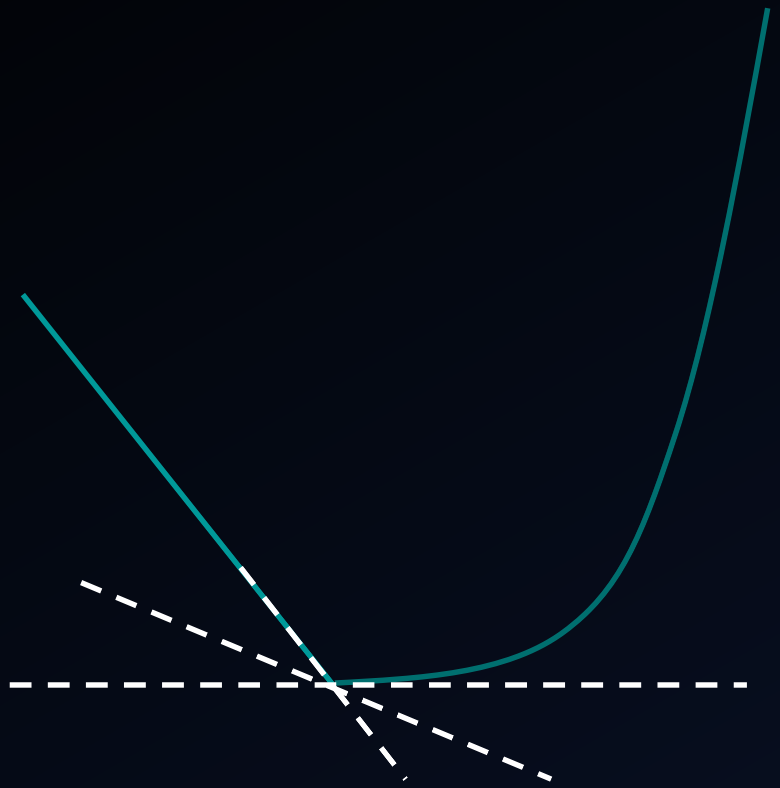
- Convex f is point-wise maximum of affine minorants
- Provides dual definition:
  - $f(x) = \max_{y \in Y} a_y^T x - b_y$, equivalently:
  - $\exists f^* \ni \; f(x) = \max_{y \in dom\, f^*} y^T x - f^*(y)$
  - f^* is called conjugate or Fenchel dual or Legendre transformation ($f^{**} = f$).
- If $f^*(y)$ is indicator of set S we get conic f:
  - $f(x) = \max_{y \in S} y^T x$

On Conjugates and Support Functions



- Convex f is point-wise maximum of affine minorants
- Provides dual definition:
  - $f(x) = \max_{y \in Y} a_y^T x - b_y$, equivalently:
  - $\exists f^* \ni \ f(x) = \max_{y \in dom\ f^*} y^T x - f^*(y)$
  - f^* is called conjugate or Fenchel dual or Legendre transformation ($f^{**} = f$).
- If $f^*(y)$ is indicator of set S we get conic f:
  - $f(x) = \max_{y \in S} y^T x$
- If S is a norm ball, we get dual norm

Connection with sub-gradient



Let,

- $y^* \in \underset{y \in dom\ f}{\operatorname{argmax}}\, y^T x - f(y)$ i.e., $f^*(x) + f(y^*) = x^T y^*$
- Then $\boldsymbol{y^*}$ **must be a sub-gradient of $\boldsymbol{f^*}$ at $\boldsymbol{x}$**
  - dual form exposes sub-gradients
- If $f^*(y)$ is indicator of set S we get conic f:
  - $f(x) = \max_{y \in S} y^T x$

Conjugates e.g.

$f(x)$	$f^*(x)$	Projection?
$\|x\|_p$	$\|x\|_{\frac{p}{p-1}}$	No $(p \notin \{1,2,\infty\})$
$\|\sigma(X)\|_p$	$\|\sigma(X)\|_{\frac{p}{p-1}}$	No $(p \notin \{1,2,\infty\})$

- $\|x\|_1$ Projection, conjugate $= O(n \log n), O(n)$
- $\|\sigma(X)\|_1$ Projection, conjugate = Full, First SVD

Rate of Convergence

Theorem[Ma11]: If X is compact convex set and f is smooth with const. L, and $s_k = \frac{2}{k+2}$, then the iterates generated by Frank-Wolfe satisfy:

$$f(x_k) - f(x^*) \leq \frac{4L\, d(X)^2}{k+2}.$$

Sub-optimal

Proof Sketch:

- $f(x_{k+1}) \leq f(x_k) + s_k \nabla f(x_k)^T (y_{k+1} - x_k) + \frac{s_k^2 L}{2} d(X)^2$
- $\Delta_{k+1} \leq (1 - s_k)\Delta_k + \frac{s_k^2 L}{2} d(X)^2$ (Solve recursion)

Rate of Convergence

Theorem[Ma11]**:** If X is compact convex set and f is smooth with const. L, and $s_k = \frac{2}{k+2}$, then the iterates generated by Frank-Wolfe satisfy:

$$f(x_k) - f(x^*) \leq \frac{4L\, d(X)^2}{k+2}.$$

Proof Sketch:

- $f(x_{k+1}) \leq f(x_k) + s_k \nabla f(x_k)^T (y_{k+1} - x_k) + \frac{s_k^2 L}{2} d(X)^2$
- $\Delta_{k+1} \leq (1 - s_k)\Delta_k + \frac{s_k^2 L}{2} d(X)^2$ (Solve recursion)

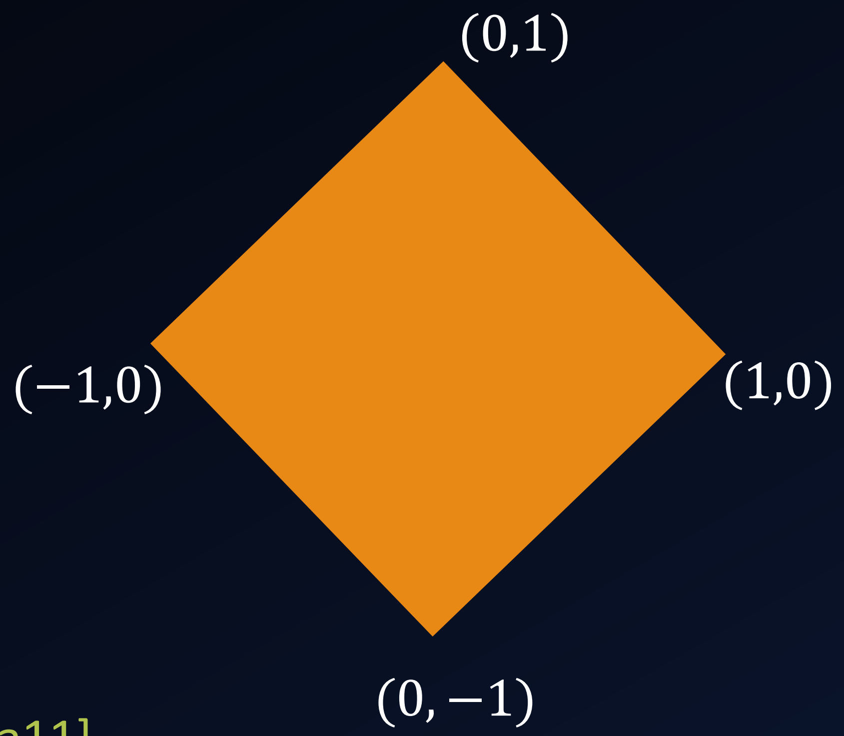
Sparse Representation – Optimality

- If $x_0 = 0$ and domain is l_1 ball, $x_k \in R^{k,n}$
  - We get exact sparsity! (unlike proj. grad.)
- Sparse representation by extreme points

- $\epsilon \approx O(L\,d(X)^2/k)$ need atleast k non-zeros [Ma11]
- Optimal in terms of accuracy-sparsity trade-off
  - Not in terms of accuracy-iterations

Sparse Representation – Optimality

- If $x_0 = 0$ and domain is $l_1$ball, $x_k \in R^{k,n}$
  - We get exact sparsity! (unlike proj. grad.)
- Sparse representation by extreme points



- $\epsilon \approx O\left({}^{L\, d(X)^2}/_{k}\right)$ need atleast k non-zeros [Ma11]
- Optimal in terms of accuracy-sparsity trade-off
  - Not in terms of accuracy-iterations

Summary comparison of always feasible methods

Property	Projected Gr.	Frank-Wolfe
Rate of convergence	+	-
Sparse Solutions	-	+
Iteration Complexity	-	+
Affine Invariance	-	+

Functional Constrained

BASED METHODS

Assumptions

$$\min_{x \in R^n} f_0(x)$$
$$s.t.\, f_i(x) \leq 0\ \forall\, i = 1, \dots, n$$

- All f_0, f_i are smooth
- L is max. const. among all

Algorithm

At iteration $1 \leq k \leq N$:

- Check if $f_i(x_k) \leq \frac{R}{\sqrt{N}} \|\nabla f_i(x_k)\| \ \forall\, i$
  - If yes, then "productive" step: $i(k) = 0$
  - If no, then "non-productive" step: $i(k)$ set to a violator
- $x_{k+1} = x_k - \frac{R}{\sqrt{N}\|\nabla f_{i(k)}(x_k)\|} \nabla f_{i(k)}(x_k)$
- Output: $\hat{x}_N$, the best among the productive.

Does it converge?

Theorem[Ju12]: Let X be bounded and L be the smoothness const. (upper bound). Then,

- $f_0(\hat{x}_N) - f_0(x^*) \leq \frac{LR}{\sqrt{N}}$
- $f_i(\hat{x}_N) \leq \frac{LR}{\sqrt{N}}\ \forall\, i$

Proof Sketch: Let $f_0(\hat{x}_N) - f_0(x^*) > \frac{LR}{\sqrt{N}}$

- $\sum_{k=1}^{N} (x_k - x^*)^T \nabla f_{i(k)}(x_k) / \|\nabla f_{i(k)}(x_k)\| \leq RN$
  - Non-productive: $\frac{R}{\sqrt{N}} (x_k - x^*)^T \nabla f_{i(k)}(x_k) / \|\nabla f_{i(k)}(x_k)\| \geq \frac{R^2}{N}$
  - Productive: $\frac{R}{\sqrt{N}} (x_k - x^*)^T \nabla f_{i(k)}(x_k) / \|\nabla f_{i(k)}(x_k)\| > \frac{R^2}{N}$

Composite Objective

PROX BASED METHODS

Composite Objectives

Non-Smooth g(w)

Smooth f(w)

$$\min_{w \in R^n} \boxed{\Omega(w)} + \boxed{\sum_{i=1}^{m} l(w'\phi(x_i), y_i)}$$

Key Idea: Do not approximate non-smooth part

Proximal Gradient Method

- $x_{k+1} = \text{argmin}_x\, f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2s_k}\|x - x_k\|^2 + g(x)$
- If g is indicator, then same as projected gr.
- If g is support function: $g(x) = \max_{y \in S} x^T y$
  - Assume min-max interchange

$$x_{k+1} = x_k - s_k \nabla f(x_k) - s_k \Pi_S\left(\frac{1}{s_k}\left(x_k - s_k \nabla f(x_k)\right)\right)$$

Proximal Gradient Method

- $x_{k+1} = \text{argmin}_x \, f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2 + g(x)$
- If g is indicator, then same as projected gr.
- If g is support function: $g(x) = \max_{y \in S} x^T y$
  - Assume min-max interchange

Again, projection

$$x_{k+1} = x_k - s_k \nabla f(x_k) - s_k \Pi_S \left(\frac{1}{s_k} (x_k - s_k \nabla f(x_k)) \right)$$

Rate of Convergence

Theorem[Ne04]: If f is smooth with const. L, and $s_k = \frac{1}{L}$, then proximal gradient method generates x_k such that:

$$f(x_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}.$$

- Can be accelerated to $O(1/k^2)$
- Composite same rate as smooth provided proximal oracle exists!

Bibliography

- **[Ne04]** Nesterov, Yurii. *Introductory lectures on convex optimization : a basic course*. Kluwer Academic Publ., 2004. http://hdl.handle.net/2078.1/116858.
- **[Ne83]** Nesterov, Yurii. *A method of solving a convex programming problem with convergence rate O (1/k2)*. Soviet Mathematics Doklady, Vol. 27(2), 372-376 pages.
- [Mo12] Moritz Hardt, Guy N. Rothblum and Rocco A. Servedio. *Private data release via learning thresholds*. SODA 2012, 168-187 pages.
- [Be09] Amir Beck and Marc Teboulle. *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*. SIAM Journal of Imaging Sciences, Vol. 2(1), 2009. 183-202 pages.
- [De13] Olivier Devolder, François Glineur and Yurii Nesterov. *First-order methods of smooth convex optimization with inexact oracle.* Mathematical Programming 2013.
- [FW59] Marguerite Frank and Philip Wolfe. An Algorithm for Quadratic Programming. Naval Research Logistics Quarterly, 1959, Vol 3, 95-110 pages.

Bibliography

- **[Ma11]** Martin Jaggi. Sparse Convex Optimization Methods for Machine Learning. PhD Thesis, 2011.
- **[Ju12]** A Juditsky and A Nemirovski. First Order Methods for Non-smooth Convex Large-Scale Optimization, I: General Purpose Methods. Optimization methods for machine learning. The MIT Press, 2012. 121-184 pages.

Thanks for listening