Frank-Wolfe Algorithm
&
Alternating Direction Method of Multipliers

Ives Macêdo
ijamj@cs.ubc.ca

October 27, 2015

Where were we?

Previous episode...

Proximal-gradient methods

$$\underset{x \in \mathcal{X}}{\text{minimize}} \quad f(x) + \varphi(x)$$

- $f : \mathcal{X} \to \mathbb{R}$ convex with Lipschitz-continuous gradient
- $\varphi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ convex and *simple* (i.e., proximable)

$$\operatorname{prox}_{\alpha\varphi} : \mathcal{X} \to \mathcal{X} \quad (\forall \alpha > 0)$$

$$\operatorname{prox}_{\alpha\varphi}(x) := \underset{\hat{x} \in \mathcal{X}}{\arg\min} \left\{ \alpha\varphi(\hat{x}) + \frac{1}{2} \|\hat{x} - x\|_2^2 \right\}$$

$$x^{k+1} := \operatorname{prox}_{\alpha_k \varphi}\left(x^k - \alpha_k \nabla f(x^k)\right)$$

Proximal-gradient methods

Good news

- $\alpha_k \equiv \alpha \in (0, 2/L) \Rightarrow f(x^k) + \varphi(x^k) - \min\{f + \varphi\} \leq O(1/k)$
- Acceleration gives $O(1/k^2)$
- Generalize projected gradient methods, where

$$\boxed{\varphi(x) = \delta_{\mathcal{C}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ +\infty & \text{if } x \notin \mathcal{C} \end{cases}}$$

Proximal-gradient methods

Bad news

- Some sets $\mathcal{C}$ can be tough to project onto but you can minimize linear functions in them

- Dealing with $\varphi(x) = \phi(Ax)$ ain't easy even when ϕ is *simple*

Proximal-gradient methods

Bad news

- Some sets $\mathcal{C}$ can be tough to project onto but you can minimize linear functions in them

Frank-Wolfe Algorithm/Conditional Gradient Method

- Dealing with $\varphi(x) = \phi(Ax)$ ain't easy even when ϕ is *simple*

Proximal-gradient methods

Bad news

- Some sets $\mathcal{C}$ can be tough to project onto but you can minimize linear functions in them

Frank-Wolfe Algorithm/Conditional Gradient Method

- Dealing with $\varphi(x) = \phi(Ax)$ ain't easy even when ϕ is *simple*

Alternating Direction Method of Multipliers (ADMM)

Frank-Wolfe Algorithm

A motivating problem

Matrix completion

$$\begin{pmatrix} ? & ? & 2 & ? \\ 1 & ? & ? & 3 \\ 1 & 2 & 2 & 3 \\ ? & 6 & 6 & 9 \\ 3 & ? & ? & 9 \\ 1 & ? & 2 & ? \\ ? & ? & 6 & 9 \end{pmatrix}$$

A motivating problem

Matrix completion

$$\begin{pmatrix} 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 3 \\ 1 & 2 & 2 & 3 \\ 0 & 6 & 6 & 9 \\ 3 & 0 & 0 & 9 \\ 1 & 0 & 2 & 0 \\ 0 & 0 & 6 & 9 \end{pmatrix}$$

A motivating problem

Matrix completion

$$\operatorname{rank}\begin{pmatrix} 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 3 \\ 1 & 2 & 2 & 3 \\ 0 & 6 & 6 & 9 \\ 3 & 0 & 0 & 9 \\ 1 & 0 & 2 & 0 \\ 0 & 0 & 6 & 9 \end{pmatrix} = 4$$

A motivating problem

Matrix completion

$$\operatorname{rank}\begin{pmatrix} 1 & 2 & 2 & 3 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 2 & 3 \\ 3 & 6 & 6 & 9 \\ 3 & 6 & 6 & 9 \\ 1 & 2 & 2 & 3 \\ 3 & 6 & 6 & 9 \end{pmatrix} = 1$$

A motivating problem

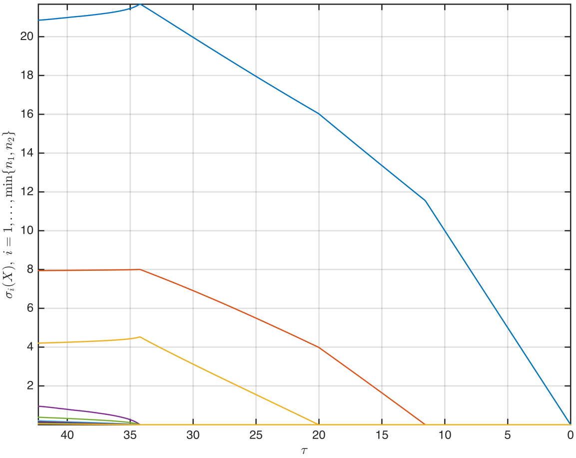
Matrix completion with nuclear-norm lasso

$$\operatorname*{minimize}_{X \in \mathbb{R}^{n_1 \times n_2}} \quad \frac{1}{2} \sum_{k=1}^{m} (X_{i_k, j_k} - b_k)^2 \quad \text{subject to} \quad \|\sigma(X)\|_1 \leq \tau$$

A motivating problem

Matrix completion with nuclear-norm lasso

$$\underset{X\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}} \quad \frac{1}{2}\sum_{k=1}^{m}(X_{i_k,j_k}-b_k)^2 \quad \text{subject to} \quad \|\sigma(X)\|_1 \le \tau$$



A motivating problem

Matrix completion with nuclear-norm lasso

$$\operatorname*{minimize}_{X \in \mathbb{R}^{n_1 \times n_2}} \quad \frac{1}{2} \sum_{k=1}^{m} (X_{i_k, j_k} - b_k)^2 \quad \text{subject to} \quad \|X\|_1 \leq \tau$$

- $\|X\|_1 = \|\sigma(X)\|_1 = \sum_{i=1}^{\min\{n_1, n_2\}} \sigma_i(X)$
- Projection onto $\{X \mid \|X\|_1 \leq \tau\}$ potentially requires full SVD
- Linear minimization requires only one SVD triplet!

Model problem

$$\underset{x \in \mathcal{C}}{\text{minimize}} \quad f(x)$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable
- $\mathcal{C} \subset \mathbb{R}^n$ is convex and compact (i.e., closed and bounded)
  - we can minimize linear functions over $\mathcal{C}$, i.e., $\forall c \in \mathbb{R}^n$

$$\text{find} \quad \hat{x} \in \underset{x \in \mathcal{C}}{\arg\min} \, \langle c, x \rangle$$

Frank-Wolfe algorithm

Frank and Wolfe (1956)

$$x^0 \in \mathcal{C}$$

$$\hat{x}^{k+1} \in \arg\min_{x \in \mathcal{C}} \Big\{ f(x^k) + \Big\langle \nabla f(x^k), x - x^k \Big\rangle \Big\}$$

$$x^{k+1} = (1 - \alpha_k) x^k + \alpha_k \hat{x}^{k+1}, \quad \alpha_k := \frac{2}{k+2}$$

Frank-Wolfe algorithm

Frank and Wolfe (1956)

$$x^0 \in \mathcal{C}$$

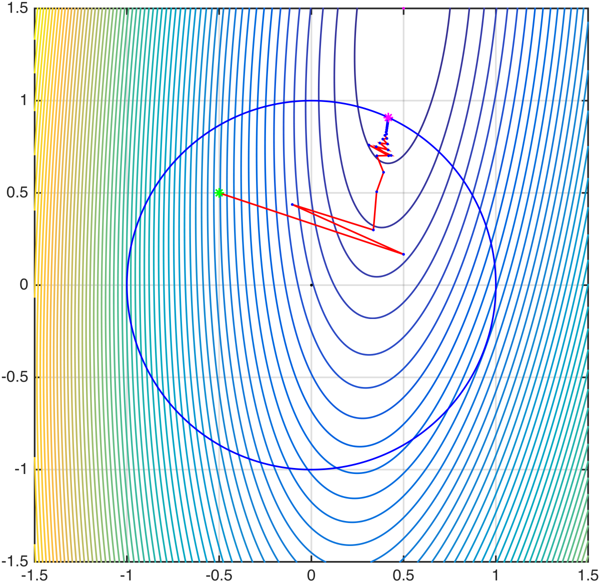
$$\hat{x}^{k+1} \in \arg\min_{x \in \mathcal{C}} \Big\{ f(x^k) + \Big\langle \nabla f(x^k), x - x^k \Big\rangle \Big\}$$

$$x^{k+1} = (1 - \alpha_k) x^k + \alpha_k \hat{x}^{k+1}, \quad \alpha_k := \frac{2}{k+2}$$

Approximation similar to projected gradient, but no quadratic term!

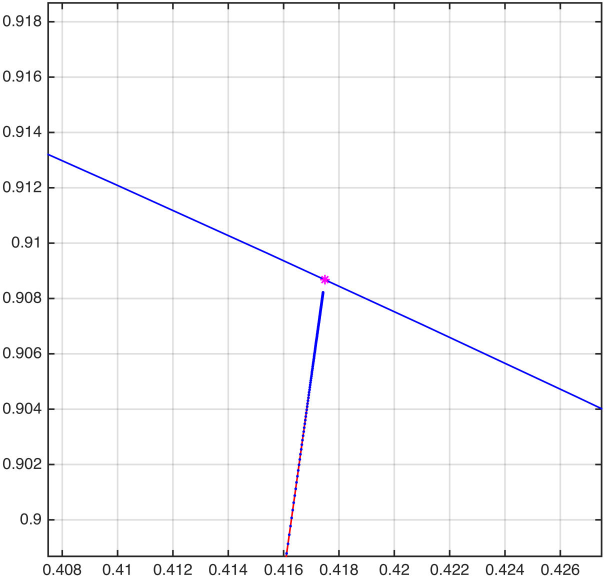
Frank-Wolfe algorithm

Visualizing the iterates



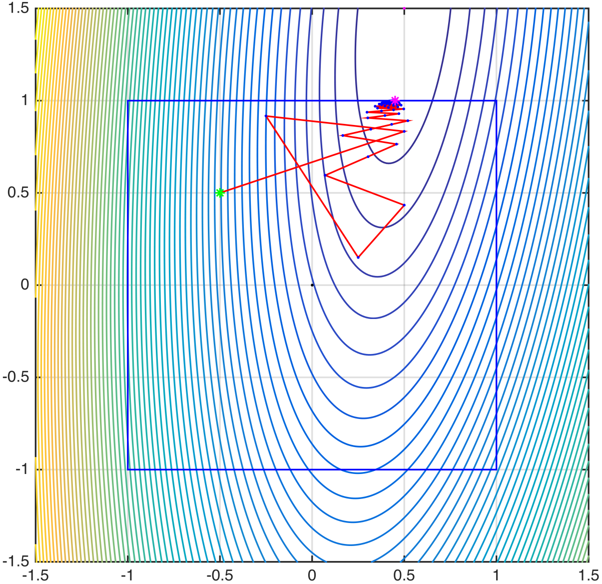
Frank-Wolfe algorithm

Visualizing the iterates



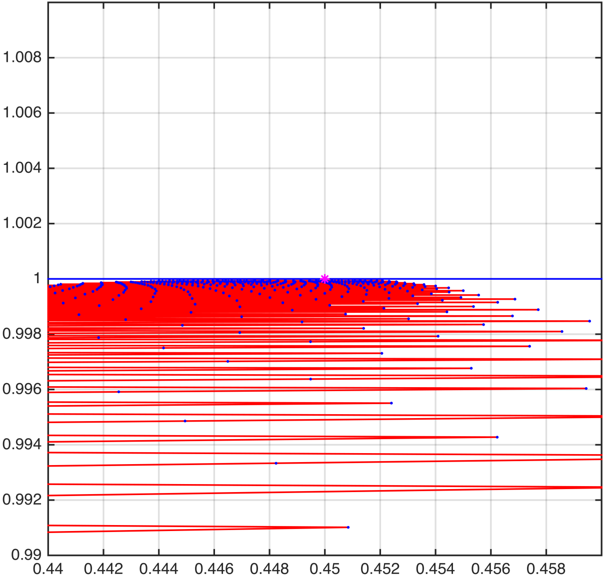
Frank-Wolfe algorithm

Visualizing the iterates



Frank-Wolfe algorithm

Visualizing the iterates



Frank-Wolfe algorithm

Curvature constant

$$\ell_f(y; x) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

$$C_f := \max_{\substack{x, \hat{x} \in \mathcal{C} \\ \alpha \in [0,1] \\ y = (1-\alpha)x + \alpha \hat{x}}} \frac{2}{\alpha^2} \ell_f(y; x)$$

Frank-Wolfe algorithm

Curvature constant (example)

$$f(x) = \frac{1}{2} \|x\|_2^2$$

$$\ell_f(y; x) = \frac{1}{2} \|y - x\|_2^2$$

$$C_f = \max_{x, \hat{x} \in \mathcal{C}} \|\hat{x} - x\|_2^2 = (\operatorname{diam} \mathcal{C})^2$$

Frank-Wolfe algorithm

Curvature constant (example)

$$f(x) = \frac{1}{2}\|x\|_2^2$$

$$\ell_f(y;x) = \frac{1}{2}\|y-x\|_2^2$$

$$C_f = \max_{x,\hat{x}\in\mathcal{C}} \|\hat{x}-x\|_2^2 = (\operatorname{diam}\mathcal{C})^2$$

If ∇f is L-Lipschitz, then $C_f \leqslant L(\operatorname{diam}\mathcal{C})^2$

Frank-Wolfe algorithm

Approximate subproblem minimizers

$$\hat{x}^{k+1} \in \left\{ \hat{x} \in \mathcal{C} \;\middle|\; \ell_f(\hat{x}; x^k) \leq \min_{x \in \mathcal{C}} \ell_f(x; x^k) + \frac{1}{2}\delta\alpha_k C_f \right\}$$

Frank-Wolfe algorithm

Exact line-search

$$\alpha_k \in \operatorname*{arg\,min}_{\alpha \in [0,1]} f\left((1-\alpha)x^k + \alpha \hat{x}^{k+1}\right)$$

Frank-Wolfe algorithm

Fully-corrective reoptimization

$$x^{k+1} \in \operatorname*{arg\,min}_{x \in \mathrm{conv}\{x^0, \hat{x}^1, \ldots, \hat{x}^{k+1}\}} f(x)$$

Frank-Wolfe algorithm

Primal-convergence

Theorem (Jaggi, 2013)

$$\boxed{f(x^k) - \inf_{\mathcal{C}} f \quad \leq \quad \frac{2C_f}{k+2}(1+\delta)}$$

Frank-Wolfe algorithm

Lower bound on primal convergence

Theorem (Canon and Cullum, 1968)

There are instances with strongly convex objectives for which the original FWA generates sequences with the following behavior: for all $C, \epsilon > 0$ there are infinitely many k such that

$$\boxed{f(x^k) - \inf_{\mathcal{C}} f \quad \geq \quad \frac{C}{k^{1+\epsilon}}}$$

Frank-Wolfe algorithm

Faster variants

- Linear convergence can be obtained in certain cases if "away/drop steps" are used; see (GuéLat and Marcotte, 1986) and (Lacoste-Julien and Jaggi, 2014)

- For smooth f and strongly convex $\mathcal{C}$, a simple variant has complexity $O(1/k^2)$ (Garber and Hazan, 2015)

Alternating Direction Method of Multipliers

A motivating problem

Total-variation image denoising

A motivating problem

Total-variation image denoising

A motivating problem

Total-variation image denoising

A motivating problem

Total-variation image denoising

A motivating problem

Total-variation image denoising

$$\boxed{\underset{f}{\text{minimize}} \quad \frac{1}{2}\int (f - f_\eta)^2 + \lambda \int \|\nabla f\|_2}$$

A motivating problem

Total-variation image denoising

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|x - x_\eta\|_2^2 + \lambda\|Dx\|_1$$

A motivating problem

Total-variation image denoising

$$\underset{x,y}{\text{minimize}} \quad \frac{1}{2}\|x - x_\eta\|_2^2 + \lambda\|y\|_1 \quad \text{subject to} \quad Dx - y = 0$$

Equality-constrained optimization

First-order optimality conditions

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad h(x) = 0$$

Necessary for $\bar{x}$ to be a minimizer:

$$\begin{aligned} \nabla f(\bar{x}) + \nabla h(\bar{x})\bar{z} &= 0 \\ h(\bar{x}) &= 0 \end{aligned}$$

Equality-constrained optimization

Quadratic penalization

$$x^{k+1} \in \arg\min_x \left\{ f(x) + \frac{\rho_k}{2} \|h(x)\|_2^2 \right\}$$

$$\nabla f(x^{k+1}) + \nabla h(x^{k+1})[\rho_k h(x^{k+1})] = 0$$

- Need $h(x^{k+1}) \to 0$ and $\rho_k \to +\infty$ for $\rho_k h(x^{k+1}) \to \bar{z} \neq 0$

Equality-constrained optimization

Lagrangian minimization

$$
\begin{aligned}
x^{k+1} &\in \arg\min_x \left\{ f(x) + \left\langle z^k, h(x) \right\rangle \right\} \\
z^{k+1} &= z^k + \alpha_k h(x^{k+1})
\end{aligned}
$$

$$\nabla f(x^{k+1}) + \nabla h(x^{k+1}) z^k = 0$$

- (Super)gradient *ascent* on concave dual
- Stability issues when argmin has multiple points at solution

Equality-constrained optimization

Method of Multipliers/Augmented Lagrangian

- MM $\approx$ Lagrangian Minimization + Quadratic Penalization

$$
\begin{aligned}
x^{k+1} &\in \arg\min_x \left\{ f(x) + \left\langle z^k, h(x) \right\rangle + \frac{\rho_k}{2} \|h(x)\|_2^2 \right\} \\
z^{k+1} &= z^k + \rho_k h(x^{k+1})
\end{aligned}
$$

$$
\nabla f(x^{k+1}) + \nabla h(x^{k+1}) z^{k+1} = 0
$$

- Will work once ρ_k *sufficiently* large (no need for $\rho_k \to +\infty$)
- Computing x^{k+1} can be tough

Model problem

$$\boxed{\underset{x \in \mathcal{X}}{\text{minimize}} \quad f(x) + \lambda \phi(Ax)}$$

- $A : \mathcal{X} \to \mathcal{Y}$ linear
- $\phi : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ convex and proximable
- $f : \mathcal{X} \to \mathbb{R}$ such that one can solve:

$$\underset{x \in \mathcal{X}}{\text{minimize}} \quad f(x) + \frac{1}{2} \|b - Ax\|_2^2$$

Model problem

Method of Multipliers

$$\underset{x,y}{\text{minimize}} \quad f(x) + \lambda\phi(y) \quad \text{subject to} \quad Ax - y = 0$$

$$(x^{k+1}, y^{k+1}) \in \arg\min_{x,y} \left\{ f(x) + \lambda\phi(y) + \frac{\rho_k}{2} \left\| Ax - y + \frac{z^k}{\rho_k} \right\|_2^2 \right\}$$

$$z^{k+1} = z^k + \rho_k(Ax^{k+1} - y^{k+1})$$

- Still tricky joint minimization over x and y
- Alternate!

Model problem

Alternating Direction Method of Multipliers

$$x^{k+1} \in \arg\min_x \left\{ f(x) + \frac{\rho_k}{2} \left\| Ax - y^k + \frac{z^k}{\rho_k} \right\|_2^2 \right\}$$

$$\begin{aligned} y^{k+1} &= \arg\min_y \left\{ \lambda\phi(y) + \frac{\rho_k}{2} \left\| Ax^{k+1} - y + \frac{z^k}{\rho_k} \right\|_2^2 \right\} \\ &= \operatorname{prox}_{\rho_k^{-1}\lambda\phi} \left[Ax^{k+1} + \frac{z^k}{\rho_k} \right] \end{aligned}$$

$$z^{k+1} = z^k + \rho_k (Ax^{k+1} - y^{k+1})$$

Model problem

Alternating Direction Method of multipliers

- Simpler iterations when $\rho_k \equiv \rho$ (defining $\hat{z}^k := z^k/\rho$)

$$x^{k+1} \in \arg\min_x \left\{ f(x) + \frac{\rho}{2} \left\| Ax - y^k + \hat{z}^k \right\|_2^2 \right\}$$

$$y^{k+1} = \operatorname{prox}_{\rho^{-1}\lambda\phi} \left[Ax^{k+1} + \hat{z}^k \right]$$

$$\hat{z}^{k+1} = \hat{z}^k + (Ax^{k+1} - y^{k+1})$$

ADMM

Total-variation denoising

ADMM

Total-variation denoising

ADMM

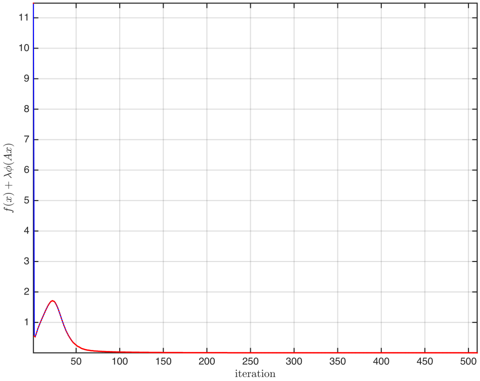
Total-variation denoising

ADMM

Convergence

- Function values decrease as $O(1/k)$ (He and Yuan, 2012)

- Linear convergence if f or ϕ is strongly convex and under certain conditions on A (Deng and Yin, 2012)



Other problems suitable for ADMM

In case I haven't bored you out of your mind...

ADMM

What if f is only proximable?

$$\underset{x}{\text{minimize}} \quad f(x) + \lambda\phi(Ax)$$

$$\begin{aligned}
&\underset{x_1,x_2,y}{\text{minimize}} \quad f(x_1) + \lambda\phi(y) \\
&\text{subject to} \\
&\qquad Ax_2 - y = 0 \\
&\qquad x_1 - x_2 = 0
\end{aligned}$$

ADMM

What if f is only proximable?

$$\begin{aligned}
&\underset{x_1, x_2, y}{\text{minimize}} \quad f(x_1) + \lambda\phi(y) \\
&\text{subject to} \\
&\qquad Ax_2 - y = 0 \\
&\qquad x_1 - x_2 = 0
\end{aligned}$$

$$\begin{aligned}
x_1^{k+1} &= \text{prox}_{\rho^{-1} f}\left[x_2^k - \hat{z}_2^k\right] \\
y^{k+1} &= \text{prox}_{\rho^{-1}\lambda\phi}\left[Ax_2^k + \hat{z}_1^k\right] \\
x_2^{k+1} &= (I + A^*A)^{-1}(x_1^{k+1} + \hat{z}_2^k + A^*(y^{k+1} - \hat{z}_1^k)) \\
\hat{z}_1^{k+1} &= \hat{z}_1^k + (Ax_2^{k+1} - y^{k+1}) \\
\hat{z}_2^{k+1} &= \hat{z}_2^k + (x_1^{k+1} - x_2^{k+1})
\end{aligned}$$

ADMM

Sum of proximable functions

$$\underset{x}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x)$$

ADMM

Sum of proximable functions

$$\underset{x, x_1, \ldots, x_m}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x_i) \quad \text{subject to} \quad x_i - x = 0, \; i = 1, \ldots, m$$

ADMM

Sum of proximable functions

$$\underset{x, x_1, \ldots, x_m}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x_i) \quad \text{subject to} \quad x_i - x = 0, \; i = 1, \ldots, m$$

$$x_i^{k+1} = \mathrm{prox}_{\rho^{-1} f_i}[x^k - \hat{z}_i^k]$$

$$x^{k+1} = \frac{1}{m} \sum_{i=1}^{m} (x_i^{k+1} + \hat{z}_i^k)$$

$$\hat{z}_i^{k+1} = \hat{z}_i^k + (x_i^{k+1} - x^{k+1})$$

ADMM

Sum of proximable functions

$$\underset{x, x_1, \ldots, x_m}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x_i) \quad \text{subject to} \quad x_i - x = 0, \; i = 1, \ldots, m$$

$$x_i^{k+1} = \mathrm{prox}_{\rho^{-1} f_i}[x^k - \hat{z}_i^k]$$

$$x^{k+1} = \frac{1}{m} \sum_{i=1}^{m} (x_i^{k+1} + \hat{z}_i^k)$$

$$\hat{z}_i^{k+1} = \hat{z}_i^k + (x_i^{k+1} - x^{k+1})$$

Distributed consensus

ADMM

Regularized sum of proximable functions

$$\underset{x}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x) + \lambda \varphi(x)$$

ADMM

Regularized sum of proximable functions

$$\underset{x,x_1,\ldots,x_m}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x_i) + \lambda\varphi(x) \quad \text{subject to} \quad x_i - x = 0,\ \forall i$$

ADMM

Regularized sum of proximable functions

$$\underset{x, x_1, \ldots, x_m}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x_i) + \lambda \varphi(x) \quad \text{subject to} \quad x_i - x = 0, \ \forall i$$

$$\begin{aligned}
x_i^{k+1} &= \operatorname{prox}_{\rho^{-1} f_i}[x^k - \hat{z}_i^k] \\
x^{k+1} &= \operatorname{prox}_{(m\rho)^{-1}\lambda\varphi}\left[\frac{1}{m}\sum_{i=1}^{m}(x_i^{k+1} + \hat{z}_i^k)\right] \\
\hat{z}_i^{k+1} &= \hat{z}_i^k + (x_i^{k+1} - x^{k+1})
\end{aligned}$$

Frank-Wolfe Algorithm
&
Alternating Direction Method of Multipliers

Ives Macêdo
ijamj@cs.ubc.ca

October 27, 2015