Primal and Dual Variables Decomposition Methods in Convex Optimization

Amir Beck

Technion - Israel Institute of Technology
Haifa, Israel

Based on joint works with
Edouard Pauwels, Shoham Sabach, Luba Tetruashvili, Yakov Vaisbourd

MIIS 2016, Chinese University of Hong Kong, Shenzhen, 18 December 2016

A (too?) General Model

$$(P) \quad \min\{H(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p) : \mathbf{x}_i \in \mathbb{R}^{n_i}\}$$

- $H : \mathbb{R}^n \to (-\infty, \infty]$ proper.
- $n = \sum_{i=1}^{p} n_i$.

At each iteration of a block variables decomposition method an operation involving only **one** of the block variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ is performed.

A (too?) General Model

$$(P) \quad \min\{H(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) : \mathbf{x}_i \in \mathbb{R}^{n_i}\}$$

- $H : \mathbb{R}^n \to (-\infty, \infty]$ proper.
- $n = \sum_{i=1}^p n_i$.

The Alternating Minimization method sequentially minimizes H w.r.t. each component in a cyclic manner.

Alternating Minimization At step k, given $\mathbf{x}^k$, the next iterate $\mathbf{x}^{k+1}$ is computed as follows:

For $i = 1 : p$

- $\mathbf{x}_1^{k+1} \in \underset{\mathbf{x}_1}{\operatorname{argmin}}\, H(\mathbf{x}_1, \mathbf{x}_2^k, \ldots, \mathbf{x}_p^k)$.
- $\mathbf{x}_2^{k+1} \in \underset{\mathbf{x}_2}{\operatorname{argmin}}\, H(\mathbf{x}_1^{k+1}, \mathbf{x}_2, \mathbf{x}_3^k, \ldots, \mathbf{x}_p^k)$.

$\vdots$

- $\mathbf{x}_p^{k+1} \in \underset{\mathbf{x}_p}{\operatorname{argmin}}\, H(\mathbf{x}_1^{k+1}, \ldots, \mathbf{x}_{p-1}^{k+1}, \mathbf{x}_p)$.

Block Descent Methods

- The AM method is just one example of a block descent method or variables decomposition method.
- Other variants replace for example the exact minimization step with some kind of a descent operator.

General Block Descent Method

For i=1:p

- $\mathbf{x}_i^{k+1} = T_i(\mathbf{x}_1^{k+1}\ldots,\mathbf{x}_{i-1}^{k+1},\mathbf{x}_i^k,\mathbf{x}_{i+1}^k,\ldots,\mathbf{x}_p^k)$.

$T_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ - a descent operator (such as one step of a minimization method)

Block Descent Methods

- The AM method is just one example of a block descent method or variables decomposition method.
- Other variants replace for example the exact minimization step with some kind of a descent operator.

General Block Descent Method

For i=1:p

- $\mathbf{x}_i^{k+1} = T_i(\mathbf{x}_1^{k+1} \ldots, \mathbf{x}_{i-1}^{k+1}, \mathbf{x}_i^k, \mathbf{x}_{i+1}^k, \ldots, \mathbf{x}_p^k)$.

$T_i : \mathbb{R}^n \rightarrow \mathbb{R}^{n_i}$ - a descent operator (such as one step of a minimization method)

- Additional variants of the method consider different index selection strategies other than cyclic (essentially cyclic,Gauss-Southwell)
- Deterministic index selection strategies can be replaced by randomized.

Example 1 of AM: IRLS - Iteratively Reweighted Least Squares

The model:

$$\text{(N)} \qquad \begin{array}{ll} \min & s(\mathbf{y}) + \sum_{i=1}^m \|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|_2 \\ \text{s.t.} & \mathbf{y} \in X, \end{array}$$

- $\mathbf{A}_i \in \mathbb{R}^{k_i \times n}, \mathbf{b}_i \in \mathbb{R}^{k_i}, i = 1, 2, \ldots, m.$
- s continuously differentiable over the closed and convex set $X \subseteq \mathbb{R}^n$.

Example 1 of AM: IRLS - Iteratively Reweighted Least Squares

The model:

$$(\text{N}) \quad \begin{array}{ll} \min & s(\mathbf{y}) + \sum_{i=1}^{m} \|\mathbf{A}_i\mathbf{y} + \mathbf{b}_i\|_2 \\ \text{s.t.} & \mathbf{y} \in X, \end{array}$$

- $\mathbf{A}_i \in \mathbb{R}^{k_i \times n}, \mathbf{b}_i \in \mathbb{R}^{k_i}, i = 1, 2, \ldots, m.$
- s continuously differentiable over the closed and convex set $X \subseteq \mathbb{R}^n$.

Examples:

- l_1-norm linear regression $\min \|\mathbf{By} - \mathbf{c}\|_1$
- Fermat-Weber problem

$$(FW) \quad \min \sum_{i=1}^{m} \omega_i \|\mathbf{y} - \mathbf{a}_i\|$$

- l_1-regularized least squares $\min \|\mathbf{By} - \mathbf{c}\|_2^2 + \lambda\|\mathbf{Dy}\|_1$..

IRLS - Iteratively Reweighted Least Squares

Wishful thinking...

Initialization: $\mathbf{y}_0 \in X$.
General Step $(k = 0, 1, \ldots)$**:**

$$\mathbf{y}_{k+1} \in \underset{\mathbf{y} \in X}{\operatorname{argmin}} \left\{ s(\mathbf{y}) + \frac{1}{2} \sum_{i=1}^{m} \frac{\|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|^2}{\|\mathbf{A}_i \mathbf{y}_k + \mathbf{b}_i\|} \right\}.$$

IRLS - Iteratively Reweighted Least Squares

Wishful thinking...

Initialization: $\mathbf{y}_0 \in X$.
General Step $(k = 0, 1, \ldots)$**:**

$$\mathbf{y}_{k+1} \in \operatorname*{argmin}_{\mathbf{y} \in X} \left\{ s(\mathbf{y}) + \frac{1}{2} \sum_{i=1}^{m} \frac{\|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|^2}{\|\mathbf{A}_i \mathbf{y}_k + \mathbf{b}_i\|} \right\}.$$

Practical method

η-IRLS

Input: $\eta > 0$ - a given parameter. **Initialization:** $\mathbf{y}_0 \in X$.
General Step $(k = 0, 1, \ldots)$**:**

$$\mathbf{y}_{k+1} \in \operatorname*{argmin}_{\mathbf{y} \in X} \left\{ s(\mathbf{y}) + \frac{1}{2} \sum_{i=1}^{m} \frac{\|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|^2}{\sqrt{\|\mathbf{A}_i \mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}} \right\}.$$

IRLS - Literature

- popular approach in robust regression (McCullagh, Nedler 83')
- applications in sparse recovery (Daubechies et al, 10')
- same as Weiszfeld's method (from 1937) for solving the Fermat-Weber problem ($\eta = 0$???)
- Convergence results are known only for very specific instances [Bissantz et. al. 08'(specific unconstrained model, asymptotic linear rate of convergence), Daubechies et al 10'(asy. linear rate, basis pursuit problem)]

- Auxiliary problem:

$$(N) \quad \begin{array}{ll} \min & h_\eta(\mathbf{y},\mathbf{z}) \equiv s(\mathbf{y}) + \frac{1}{2}\sum_{i=1}^m \left(\frac{\|\mathbf{A}_i\mathbf{y}+\mathbf{b}_i\|^2+\eta^2}{z_i} + z_i \right) \\ \text{s.t.} & \mathbf{y} \in X \\ & \mathbf{z} \in [\eta/2, \infty)^m, \end{array} .$$

- Minimizing w.r.t. $\mathbf{z}$ implies that the problem is a smoothed version of (N):

$$(N_\eta) \quad \begin{array}{ll} \min & s(\mathbf{y}) + \sum_{i=1}^m \sqrt{\|\mathbf{A}_i\mathbf{y}+\mathbf{b}_i\|_2^2 + \eta^2} \\ \text{s.t.} & \mathbf{y} \in X \end{array}$$

IRLS $\Leftrightarrow$ AM

- $(\mathbf{y}_k, \mathbf{z}_k)$ - the k-th iterate of the AM method.
- The $\mathbf{z}$-step in AM: $z_i = \sqrt{\|\mathbf{A}_i \mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}$

IRLS $\Leftrightarrow$ AM

- $(\mathbf{y}_k, \mathbf{z}_k)$ - the k-th iterate of the AM method.
- The **z**-step in AM: $z_i = \sqrt{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}$
- The **y**-step in AM:

$$\mathbf{y}_{k+1} \in \operatorname*{argmin}_{\mathbf{y}\in X} \left\{ s(\mathbf{y}) + \frac{1}{2}\sum_{i=1}^{m} \frac{\|\mathbf{A}_i\mathbf{y} + \mathbf{b}_i\|^2}{\sqrt{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}} \right\}.$$

IRLS $\Leftrightarrow$ AM

- $(\mathbf{y}_k, \mathbf{z}_k)$ - the k-th iterate of the AM method.
- The **z**-step in AM: $z_i = \sqrt{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}$
- The **y**-step in AM:

$$\mathbf{y}_{k+1} \in \operatorname*{argmin}_{\mathbf{y}\in X} \left\{ s(\mathbf{y}) + \frac{1}{2}\sum_{i=1}^{m} \frac{\|\mathbf{A}_i\mathbf{y} + \mathbf{b}_i\|^2}{\sqrt{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}} \right\}.$$

- The methods are equivalent given that the initial $\mathbf{z}_0$ is given by

$$[\mathbf{z}_0]_i = \sqrt{\|\mathbf{A}_i\mathbf{y}_0 + \mathbf{b}_i\|^2 + \eta^2}$$

Example 2 of AM: A Composite Model

$$T^* = \min \{ T(\mathbf{y}) \equiv q(\mathbf{y}) + r(\mathbf{A}\mathbf{y}) \},$$

- $q : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ closed, proper, convex.
- $r : \mathbb{R}^m \to \mathbb{R}$ real-valued convex function.

Example 2 of AM: A Composite Model

$$T^* = \min \{ T(\mathbf{y}) \equiv q(\mathbf{y}) + r(\mathbf{A}\mathbf{y}) \},$$

- $q : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ closed, proper, convex.
- $r : \mathbb{R}^m \to \mathbb{R}$ real-valued convex function.

- A popular penalized approach: Consider the problem

$$\min_{\mathbf{z},\mathbf{y}} \{ q(\mathbf{y}) + r(\mathbf{z}) : \mathbf{z} = \mathbf{A}\mathbf{y} \}.$$

- Write a penalized version:

$$(C) \quad T^*_\rho = \min_{\mathbf{y},\mathbf{z}} \left\{ T_\rho(\mathbf{y}, \mathbf{z}) = q(\mathbf{y}) + r(\mathbf{z}) + \frac{\rho}{2} \|\mathbf{z} - \mathbf{A}\mathbf{y}\|^2 \right\}$$

- Employ the AM method.

AM for solving (C)

Alternating Minimization for Solving (C)
Input: $\rho > 0$ - a given parameter.
Initialization: $\mathbf{y}_0 \in \mathbb{R}^n, \mathbf{z}_0 \in \operatorname{argmin}\left\{r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{z} - \mathbf{A}\mathbf{y}_0\|^2\right\}$.
General Step (k=0,1,...):

$$\begin{aligned} \mathbf{y}_{k+1} &\in \underset{\mathbf{y}\in\mathbb{R}^n}{\operatorname{argmin}}\left\{q(\mathbf{y}) + \frac{\rho}{2}\|\mathbf{z}_k - \mathbf{A}\mathbf{y}\|^2\right\}, \\ \mathbf{z}_{k+1} &= \underset{\mathbf{z}\in\mathbb{R}^m}{\operatorname{argmin}}\left\{r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{z} - \mathbf{A}\mathbf{y}_{k+1}\|^2\right\}. \end{aligned}$$

AM for solving (C)

Alternating Minimization for Solving (C)
Input: $\rho > 0$ - a given parameter.
Initialization: $\mathbf{y}_0 \in \mathbb{R}^n, \mathbf{z}_0 \in \operatorname{argmin} \left\{ r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{z} - \mathbf{A}\mathbf{y}_0\|^2 \right\}$.
General Step (k=0,1,...):

$$\begin{aligned} \mathbf{y}_{k+1} &\in \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n} \left\{ q(\mathbf{y}) + \frac{\rho}{2}\|\mathbf{z}_k - \mathbf{A}\mathbf{y}\|^2 \right\}, \\ \mathbf{z}_{k+1} &= \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^m} \left\{ r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{z} - \mathbf{A}\mathbf{y}_{k+1}\|^2 \right\}. \end{aligned}$$

Implementable in several important examples, e.g.,

$$\min \|\mathbf{C}\mathbf{x} - \mathbf{d}\|_2^2 + \|\mathbf{L}\mathbf{x}\|_1.$$

(prox of l_1+solution of a linear system)

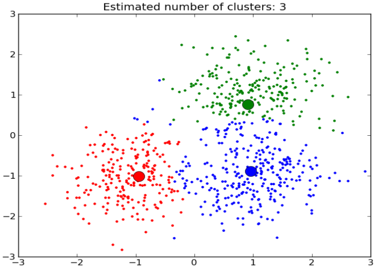
Example 3 of AM: k-means method in clustering

Input:

- n points $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\} \subseteq \mathbb{R}^d$.
- k - number of clusters .

Clusters:

- The idea is to partition the data $\mathcal{A}$ into k subsets (clusters) $\mathcal{A}_1, \ldots, \mathcal{A}_k$, called clusters.
- For each $l \in \{1, \ldots, k\}$, the cluster $\mathcal{A}_l$ is represented by its so-called center $\mathbf{x}_l$.
- The clustering problem: determine k cluster centers $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ such that the sum of distances from each point $\mathbf{a}_i$ to a nearest cluster center $\mathbf{x}_l$ is minimized.



$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_k} \sum_{i=1}^{n} \min_{l=1,2,\ldots,k} \|\mathbf{a}_i - \mathbf{x}_l\|^2.$$

Clustering: AM=k-means

Using the trick:

$$\min\{b_1, \ldots, b_k\} = \min\{\mathbf{b}^T\mathbf{y} : \mathbf{y} \in \Delta_k\}.$$

where $\Delta_k = \{\mathbf{y} \in \mathbb{R}^k : \mathbf{e}^T\mathbf{y} = 1, \mathbf{y} \geq 0\}$, we can reformulate:

$$\begin{array}{ll} \min & \sum_{i=1}^n \sum_{l=1}^k y_l^i \|\mathbf{a}_i - \mathbf{x}_l\|^2 \\ \text{s.t.} & \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d, \\ & \mathbf{y}^1, \ldots, \mathbf{y}^n \in \Delta_k, \end{array}$$

Clustering: AM=k-means

Using the trick:

$$\min\{b_1, \ldots, b_k\} = \min\{\mathbf{b}^T\mathbf{y} : \mathbf{y} \in \Delta_k\}.$$

where $\Delta_k = \{\mathbf{y} \in \mathbb{R}^k : \mathbf{e}^T\mathbf{y} = 1, \mathbf{y} \geq 0\}$, we can reformulate:

$$\begin{array}{ll} \min & \sum_{i=1}^n \sum_{l=1}^k y_l^i \|\mathbf{a}_i - \mathbf{x}_l\|^2 \\ \text{s.t.} & \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d, \\ & \mathbf{y}^1, \ldots, \mathbf{y}^n \in \Delta_k, \end{array}$$

k-**means**. repeat:

- **Assignment step.** assign each point $\mathbf{a}_i$ to closest cluster center:

$$\mathcal{A}_l = \{i : \|\mathbf{a}_i - \mathbf{x}_l\| \leq \|\mathbf{a}_i - \mathbf{x}_j\| \text{ for all } j = 1, \ldots, k\}, l = 1, 2, \ldots, k.$$

- **Update step.** Cluster centers are averages: $\mathbf{x}_l = \frac{1}{|\mathcal{A}_l|} \sum_{i \in \mathcal{A}_l} \mathbf{a}_i$.

Example 4 of AM: proximal point method

Consider the model:

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad (f : \mathbb{R}^n \to (-\infty, \infty])$$

Example 4 of AM: proximal point method

Consider the model:

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad (f : \mathbb{R}^n \to (-\infty, \infty])$$

Rewrite the problem as follows:

$$\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + \frac{c}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

Example 4 of AM: proximal point method

Consider the model:

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad (f : \mathbb{R}^n \to (-\infty, \infty])$$

Rewrite the problem as follows:

$$\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + \frac{c}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

- Same problem since minimizing w.r.t. $\mathbf{y}$ yields $\mathbf{y} = \mathbf{x}$.
- The alternating minimization method is the proximal point method:

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}) + \frac{c}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 \right\}$$

Example of Inexact Block Method: RSTLS

- **Linear inverse problem:** Given an approximate linear system

$$\mathbf{Ax} \approx \mathbf{b}$$

  find a "good" estimate of $\mathbf{x}$.

Example of Inexact Block Method: RSTLS

- **Linear inverse problem:** Given an approximate linear system

$$\mathbf{Ax} \approx \mathbf{b}$$

find a "good" estimate of $\mathbf{x}$.

- **Least squares:**

$$\hat{\mathbf{x}}_{LS} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{Ax} - \mathbf{b}\|^2 = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

assumes $\mathbf{A}$ full column rank, requires regularization?

Example of Inexact Block Method: RSTLS

- **Linear inverse problem:** Given an approximate linear system

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}$$

find a "good" estimate of $\mathbf{x}$.

- **Least squares:**

$$\hat{\mathbf{x}}_{LS} = \operatorname*{argmin}_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

assumes $\mathbf{A}$ full column rank, requires regularization?

- **From LS to TLS: A** unknown

Example of Inexact Block Method: RSTLS

- **Linear inverse problem:** Given an approximate linear system

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}$$

find a "good" estimate of $\mathbf{x}$.

- **Least squares:**

$$\hat{\mathbf{x}}_{LS} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

assumes $\mathbf{A}$ full column rank, requires regularization?

- **From LS to TLS: A** unknown

Least Squares (LS)	Total Least Squares (TLS)
$\min_{\mathbf{w},\mathbf{x}} \|\mathbf{w}\|^2$ s.t. $\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{w}$	$\min_{\mathbf{w},\mathbf{E},\mathbf{x}} \|\mathbf{E}\|^2 + \|\mathbf{w}\|^2$ s.t. $(\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b} + \mathbf{w}$
minimal perturbation to rhs which makes this linear system consistent	minimal perturbation to both rhs and lhs matrix which makes the system consistent (Golub, Van Loan (80))

From TLS to RSTLS

- **The total least squares (TLS) problem:**

$$\min_{\mathbf{x},\mathbf{E}} \|(\mathbf{A}+\mathbf{E})\mathbf{x}-\mathbf{b}\|^2 + \|\mathbf{E}\|^2$$

From TLS to RSTLS

- **The total least squares (TLS) problem:**

$$\min_{\mathbf{x},\mathbf{E}} \|(\mathbf{A}+\mathbf{E})\mathbf{x}-\mathbf{b}\|^2 + \|\mathbf{E}\|^2$$

- **The structured TLS (STLS) problem: E** has some linear structure – $\mathbf{E} = \sum_{i=1}^{p} y_i \mathbf{E}_i$

$$\min_{\mathbf{x},\mathbf{y}} \|(\mathbf{A}+\sum_{i=1}^{p} y_i \mathbf{E}_i)\mathbf{x}-\mathbf{b}\|^2 + \|\mathbf{D}\mathbf{y}\|^2$$

From TLS to RSTLS

- **The total least squares (TLS) problem:**

$$\min_{\mathbf{x},\mathbf{E}} \|(\mathbf{A}+\mathbf{E})\mathbf{x}-\mathbf{b}\|^2+\|\mathbf{E}\|^2$$

- **The structured TLS (STLS) problem:** $\mathbf{E}$ has some linear structure – $\mathbf{E}=\sum_{i=1}^{p} y_i\mathbf{E}_i$

$$\min_{\mathbf{x},\mathbf{y}} \|(\mathbf{A}+\sum_{i=1}^{p} y_i\mathbf{E}_i)\mathbf{x}-\mathbf{b}\|^2+\|\mathbf{D}\mathbf{y}\|^2$$

- **The Regularized STLS (RSTLS) problem:** regularize $\mathbf{x}$

$$\min_{\mathbf{x},\mathbf{y}} \|(\mathbf{A}+\sum_{i=1}^{p} y_i\mathbf{E}_i)\mathbf{x}-\mathbf{b}\|^2+\|\mathbf{D}\mathbf{y}\|^2+g(\mathbf{x})$$

where g is extended real-valued (can also account for constraints)

From TLS to RSTLS

- **The total least squares (TLS) problem:**

$$\min_{\mathbf{x},\mathbf{E}} \|(\mathbf{A}+\mathbf{E})\mathbf{x}-\mathbf{b}\|^2 + \|\mathbf{E}\|^2$$

- **The structured TLS (STLS) problem:** $\mathbf{E}$ has some linear structure – $\mathbf{E} = \sum_{i=1}^p y_i \mathbf{E}_i$

$$\min_{\mathbf{x},\mathbf{y}} \|(\mathbf{A}+\sum_{i=1}^p y_i\mathbf{E}_i)\mathbf{x}-\mathbf{b}\|^2 + \|\mathbf{D}\mathbf{y}\|^2$$

- **The Regularized STLS (RSTLS) problem:** regularize $\mathbf{x}$

$$\min_{\mathbf{x},\mathbf{y}} \|(\mathbf{A}+\sum_{i=1}^p y_i\mathbf{E}_i)\mathbf{x}-\mathbf{b}\|^2 + \|\mathbf{D}\mathbf{y}\|^2 + g(\mathbf{x})$$

where g is extended real-valued (can also account for constraints)

A schematic block descent method on x and y:

$$\begin{aligned}
\mathbf{y}^{k+1} &= \operatorname*{argmin}_{\mathbf{y}} \|(\mathbf{A}+\textstyle\sum_{i=1}^p y_i\mathbf{E}_i)\mathbf{x}^k-\mathbf{b}\|^2 + \|\mathbf{D}\mathbf{y}\|^2\\
\mathbf{x}^{k+1} &\approx \operatorname*{argmin}_{\mathbf{x}} \|(\mathbf{A}+\textstyle\sum_{i=1}^p y_i^{k+1}\mathbf{E}_i)\mathbf{x}-\mathbf{b}\|^2 + g(\mathbf{x}).
\end{aligned}$$

Problem in **y** - solution of a (small?) linear system. Problem in **y** - approximate "solution" of RLS (smooth+nonsmooth?)

Why Block Descent Methods?

- Simple and cheap updates at each iteration - suitable for large-scale applications.
- Allow larger step-sizes at each iteration.
- In some nonconvex settings - results with better quality solutions.

Convergence?? well...

- Convergence of the AM method is not always guaranteed.
- Powell's example (73'):

$$\begin{aligned}\varphi(x,y,z) &= -xy - yz - zx + [x-1]_+^2 + [-x-1]_+^2 \\ &\quad + [y-1]_+^2 + [-y-1]_+^2 + [z-1]_+^2 + [-z-1]_+^2.\end{aligned}$$

differentiable and nonconvex

- Fixing y, z, it is easy to show that that

$$\operatorname*{argmin}_x \varphi(x,y,z) = \begin{cases} \mathrm{sgn}(y+z)(1+\frac{1}{2}|y+z|) & y+z \neq 0 \\ [-1,1] & y+z = 0 \end{cases}$$

Similar formulas for minimizing w.r.t. y and z.

Powell's Example

- Start with $\left(-1-\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$.

First six iterations

$$\left(1+\frac{1}{8}\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$$
$$\left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$$
$$\left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$

Powell's Example

- Start with $\left(-1-\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$.

First six iterations

$$\left(1+\frac{1}{8}\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$$
$$\left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$$
$$\left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$

$$\left(-1-\frac{1}{64}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$
$$\left(-1-\frac{1}{64}\varepsilon, 1+\frac{1}{128}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$
$$\left(-1-\frac{1}{64}\varepsilon, 1+\frac{1}{128}\varepsilon, -1-\frac{1}{256}\varepsilon\right)$$

Powell's Example

- Start with $\left(-1-\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$.

First six iterations

$$\left(1+\frac{1}{8}\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$$
$$\left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$$
$$\left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$

$$\left(-1-\frac{1}{64}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$
$$\left(-1-\frac{1}{64}\varepsilon, 1+\frac{1}{128}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$
$$\left(-1-\frac{1}{64}\varepsilon, 1+\frac{1}{128}\varepsilon, -1-\frac{1}{256}\varepsilon\right)$$

- We are essentially back at the first point, but with $\frac{1}{64}\varepsilon$ instead of ε.
- The process will continue to cycle around the 6 non-stationary points

$$(1,1,-1),(1,-1,-1),(1,-1,1),(-1,-1,1),(-1,1,1),(-1,1,-1).$$

Powell's Example

- Start with $\left(-1-\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$.

First six iterations

$$\left(1+\frac{1}{8}\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$$
$$\left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$$
$$\left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$

$$\left(-1-\frac{1}{64}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$
$$\left(-1-\frac{1}{64}\varepsilon, 1+\frac{1}{128}\varepsilon, 1+\frac{1}{32}\varepsilon\right)$$
$$\left(-1-\frac{1}{64}\varepsilon, 1+\frac{1}{128}\varepsilon, -1-\frac{1}{256}\varepsilon\right)$$

- We are essentially back at the first point, but with $\frac{1}{64}\varepsilon$ instead of ε.
- The process will continue to cycle around the 6 non-stationary points

$$(1,1,-1),(1,-1,-1),(1,-1,1),(-1,-1,1),(-1,1,1),(-1,1,-1).$$

- Can be rectified if certain uniqueness/convexity/boundedness assumptions are made [Zadeh '70, Grippo Sciandrone '00, Bertsekas & Tsitsikls '89, Luo & Tseng '93,....]

Typical Convergence Result of AM

- $\bar{\mathbf{x}} \in \text{dom}(H)$ is a coordinate-wise minimum if for any i,

$$\bar{\mathbf{x}}_i \in \underset{\mathbf{x}_i}{\operatorname{argmin}}\, H(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{i-1}, \mathbf{x}_i, \bar{\mathbf{x}}_{i+1}, \ldots, \bar{\mathbf{x}}_p)$$

Typical Convergence Result of AM

- $\bar{\mathbf{x}} \in \text{dom}(H)$ is a coordinate-wise minimum if for any i,

$$\bar{\mathbf{x}}_i \in \operatorname*{argmin}_{\mathbf{x}_i} H(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{i-1}, \mathbf{x}_i, \bar{\mathbf{x}}_{i+1}, \ldots, \bar{\mathbf{x}}_p)$$

Theorem (e.g., [Bertsekas, '99]) If

- H proper, closed, continuous over its domain;
- for each $\bar{\mathbf{x}} \in \text{dom}(H)$ and i, the problem $\min_{\mathbf{x}_i} H(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{i-1}, \mathbf{x}_i, \bar{\mathbf{x}}_{i+1}, \ldots, \bar{\mathbf{x}}_p)$ attains a unique minimizer;
- level sets of H are bounded,

Then the sequence generated by the AM method is bounded and its limit points are coordinate-wise minima.

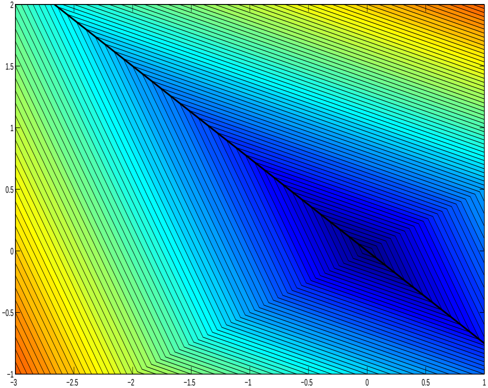
The Failure of Convexity

- Unfortunately, in the absence of differentiability, even convexity is not enough to guarantee convergence to an optimal or even stationary point.

Example:

$f(x_1, x_2) = |3x_1 + 4x_2| + |-x_1 + 2x_2|$

- All the points on the emphasized line $\{(-4\alpha, 3\alpha) : \alpha \in \mathbb{R}\}$ are coordinate-wise minima, and only (0,0) is a global minimum.



Any block descent method might converge to the non-global solution.

The Composite Model

smooth+separable convex

The Composite Model

smooth+separable convex

The composite model

$$(P) \quad \underbrace{\min f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) + \overbrace{\sum_{i=1}^{p} g_i(\mathbf{x}_i)}^{g(\mathbf{x})}}_{H(\mathbf{x}_1,\ldots,\mathbf{x}_p)}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ - continuously differentiable.
- $g_i : \mathbb{R}^{n_i} \to (-\infty, \infty]$ - closed, proper, convex.
- H with bounded level sets.

Main property: A coordinate-wise minimum of (P) is a stationary point.

$$-\nabla_i f(\mathbf{x}) \in \partial g_i(\mathbf{x}), \quad i = 1, 2, \ldots, p.$$

The Composite Model

smooth+separable convex

The composite model

$$(P) \quad \underbrace{\min f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) + \overbrace{\sum_{i=1}^{p} g_i(\mathbf{x}_i)}^{g(\mathbf{x})}}_{H(\mathbf{x}_1, \ldots, \mathbf{x}_p)}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ - continuously differentiable.
- $g_i : \mathbb{R}^{n_i} \to (-\infty, \infty]$ - closed, proper, convex.
- H with bounded level sets.

Main property: A coordinate-wise minimum of (P) is a stationary point.

$$-\nabla_i f(\mathbf{x}) \in \partial g_i(\mathbf{x}), \quad i = 1, 2, \ldots, p.$$

Corollary: Under uniqueness, limit points of AM are stationary points.

The Composite Model

smooth+separable convex

The composite model

$$(P) \quad \min \underbrace{f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) + \overbrace{\sum_{i=1}^{p} g_i(\mathbf{x}_i)}^{g(\mathbf{x})}}_{H(\mathbf{x}_1,\ldots,\mathbf{x}_p)}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ - continuously differentiable.
- $g_i : \mathbb{R}^{n_i} \to (-\infty, \infty]$ - closed, proper, convex.
- H with bounded level sets.

Main property: A coordinate-wise minimum of (P) is a stationary point.

$$-\nabla_i f(\mathbf{x}) \in \partial g_i(\mathbf{x}), \quad i = 1, 2, \ldots, p.$$

Corollary: Under uniqueness, limit points of AM are stationary points.
Theorem [generalization of Grippo and Sciandrone '00]: convergence to optimal points is guaranteed if uniqueness is replaced by convexity of f.

Examples

- **Clustering:**

$$\begin{array}{ll} \min & \sum_{i=1}^{n} \sum_{l=1}^{k} y_l^i \|\mathbf{a}_i - \mathbf{x}_l\|^2 \\ \text{s.t.} & \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d, \\ & \mathbf{y}^1, \ldots, \mathbf{y}^n \in \Delta_k, \end{array}$$

Here:

$$f(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \sum_{l=1}^{k} y_l^i \|\mathbf{a}_i - \mathbf{x}_l\|^2, g_1(\mathbf{x}) \equiv 0, g_2(\mathbf{y}) = \sum_{i=1}^{n} \delta_{\Delta_k}(\mathbf{y}^i)$$

Examples

- **Clustering:**

$$\begin{array}{ll} \min & \sum_{i=1}^{n}\sum_{l=1}^{k} y_l^i \|\mathbf{a}_i - \mathbf{x}_l\|^2 \\ \text{s.t.} & \mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^d, \\ & \mathbf{y}^1,\ldots,\mathbf{y}^n \in \Delta_k, \end{array}$$

  Here:

$$f(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\sum_{l=1}^{k} y_l^i \|\mathbf{a}_i - \mathbf{x}_l\|^2, g_1(\mathbf{x}) \equiv 0, g_2(\mathbf{y}) = \sum_{i=1}^{n} \delta_{\Delta_k}(\mathbf{y}^i)$$

- **RSTLS:**

$$\min_{\mathbf{x},\mathbf{y}} \|(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{E}_i)\mathbf{x} - \mathbf{b}\|^2 + \|\mathbf{D}\mathbf{y}\|^2 + g(\mathbf{x})$$

  Here:

$$f(\mathbf{x},\mathbf{y}) = \|(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{E}_i)\mathbf{x} - \mathbf{b}\|^2 + \|\mathbf{D}\mathbf{y}\|^2, g_1(\mathbf{x}) = g(\mathbf{x}), g_2(\mathbf{y}) \equiv 0$$

Types of Steps in Block Descent Methods

General Block Descent Method

- pick an index $i_k \in \{1, 2, \ldots, p\}$
- $\mathbf{x}_{i_k}^{k+1} = T_{i_k}(\mathbf{x}_1^k \ldots, \mathbf{x}_{i_k-1}^k, \mathbf{x}_{i_k}^k, \mathbf{x}_{i_k+1}^k, \ldots, \mathbf{x}_p^k)$, $\mathbf{x}_j^{k+1} = \mathbf{x}_j^k$, $j \neq i_k$.

Types of Steps in Block Descent Methods

General Block Descent Method

- pick an index $i_k \in \{1, 2, \ldots, p\}$
- $\mathbf{x}_{i_k}^{k+1} = T_{i_k}(\mathbf{x}_1^k \ldots, \mathbf{x}_{i_k-1}^k, \mathbf{x}_{i_k}^k, \mathbf{x}_{i_k+1}^k, \ldots, \mathbf{x}_p^k)$, $\mathbf{x}_j^{k+1} = \mathbf{x}_j^k$, $j \neq i_k$.

Two methods for solving the composite model (f smooth, g convex)

$$\min\{f(\mathbf{x}) + g(\mathbf{x})\}$$

1. **Conditional Gradient** (linearize)

$$\begin{aligned} \mathbf{p}(\mathbf{x}^k) &\in \operatorname*{argmin}_{\mathbf{p}} \left\{\langle \nabla f(\mathbf{x}^k), \mathbf{p}\rangle + g(\mathbf{p})\right\} \\ \mathbf{x}^{k+1} &= \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k)\ (t_k \in [0,1]) \end{aligned}$$

Types of Steps in Block Descent Methods

General Block Descent Method

- pick an index $i_k \in \{1, 2, \ldots, p\}$
- $\mathbf{x}_{i_k}^{k+1} = T_{i_k}(\mathbf{x}_1^k \ldots, \mathbf{x}_{i_k-1}^k, \mathbf{x}_{i_k}^k, \mathbf{x}_{i_k+1}^k, \ldots, \mathbf{x}_p^k)$, $\mathbf{x}_j^{k+1} = \mathbf{x}_j^k$, $j \neq i_k$.

Two methods for solving the composite model (f smooth, g convex)

$$\min\{f(\mathbf{x}) + g(\mathbf{x})\}$$

1. **Conditional Gradient** (linearize)

$$\begin{aligned} \mathbf{p}(\mathbf{x}^k) &\in \underset{\mathbf{p}}{\operatorname{argmin}} \left\{\langle \nabla f(\mathbf{x}^k), \mathbf{p}\rangle + g(\mathbf{p})\right\} \\ \mathbf{x}^{k+1} &= \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k)\ (t_k \in [0,1]) \end{aligned}$$

2. **Proximal Gradient** (linearize and Regularize)

$$\mathbf{x}^{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \langle \nabla f(\mathbf{x}^k), \mathbf{x}\rangle + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}^k\|^2 + g(\mathbf{x}) \right\}$$

or $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k))$

Types of Steps in Block Descent Methods

Moreau's proximal mapping:

$$\mathrm{prox}_h(\mathbf{x}) = \operatorname{argmin}\left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2 \right\}.$$

Block Proximal Gradient

$$\begin{aligned} \mathbf{x}_{i_k}^{k+1} &= \mathrm{prox}_{t_k g_{i_k}}\left(\mathbf{x}_{i_k}^k - t_k \nabla_{i_k} f(\mathbf{x}^k)\right) \\ \mathbf{x}_j^{k+1} &= \mathbf{x}_j^k, \quad j \neq i_k \end{aligned}$$

$g_i \equiv 0$ - block gradient descent, $g_i = \delta_{X_i}$ - block projected gradient.

Types of Steps in Block Descent Methods

Moreau's proximal mapping:

$$\mathrm{prox}_h(\mathbf{x}) = \operatorname{argmin}\left\{h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\right\}.$$

Block Proximal Gradient

$$\begin{aligned}\mathbf{x}_{i_k}^{k+1} &= \mathrm{prox}_{t_k g_{i_k}}\left(\mathbf{x}_{i_k}^k - t_k \nabla_{i_k} f(\mathbf{x}^k)\right)\\ \mathbf{x}_j^{k+1} &= \mathbf{x}_j^k, \quad j \neq i_k\end{aligned}$$

$g_i \equiv 0$ - block gradient descent, $g_i = \delta_{X_i}$ - block projected gradient.

Block Conditional Gradient

$$\begin{aligned}\mathbf{p}_{i_k}^k &\in \mathrm{argmin}_{\mathbf{p}_{i_k} \in \mathrm{dom} g_{i_k}}\left\{\langle \nabla_{i_k} f(\mathbf{x}^k), \mathbf{p}_{i_k}\rangle + g_{i_k}(\mathbf{p}_{i_k})\right\},\\ \mathbf{x}_{i_k}^{k+1} &= \mathbf{x}_{i_k}^k + t_k(\mathbf{p}_{i_k}^k - \mathbf{x}_{i_k}^k)\\ \mathbf{x}_j^{k+1} &= \mathbf{x}_j^k, \quad j \neq i_k\end{aligned}$$

Rates of Convergence – the Non-Block Case

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Proximal gradient $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k))$

g extended real-valued proper closed and convex

Rates of Convergence – the Non-Block Case

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Proximal gradient $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k))$

g extended real-valued proper closed and convex

- (f convex $C^{1,1}$) [B. Teboulle 09']

$$H(\mathbf{x}^k) - H^* = O(1/k)$$

Rates of Convergence – the Non-Block Case

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Proximal gradient $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k))$

g extended real-valued proper closed and convex

- (f convex $C^{1,1}$) [B. Teboulle 09']

$$H(\mathbf{x}^k) - H^* = O(1/k)$$

- (f convex $C^{1,1}$) [B. Teboulle 09']

$$H(\mathbf{x}^k) - H^* = O(1/k^2)$$

accelerated version (FISTA) $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{y}^k - t_k \nabla f(\mathbf{y}^k))$. Other multi-step methods exist (Tseng '10, Nesterov '13)

Rates of Convergence – the Non-Block Case

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Proximal gradient $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k))$

g extended real-valued proper closed and convex

- (f convex $C^{1,1}$) [B. Teboulle 09']

$$H(\mathbf{x}^k) - H^* = O(1/k)$$

- (f convex $C^{1,1}$) [B. Teboulle 09']

$$H(\mathbf{x}^k) - H^* = O(1/k^2)$$

  accelerated version (FISTA) $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{y}^k - t_k \nabla f(\mathbf{y}^k))$. Other multi-step methods exist (Tseng '10, Nesterov '13)

- (f strongly convex $C^{1,1}$)

$$H(\mathbf{x}^k) - H^* = O(q^k), q \in (0,1)$$

Rates of Convergence – the Non-Block Case

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Proximal gradient $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k))$

g extended real-valued proper closed and convex

- (f convex $C^{1,1}$) [B. Teboulle 09']

$$H(\mathbf{x}^k) - H^* = O(1/k)$$

- (f convex $C^{1,1}$) [B. Teboulle 09']

$$H(\mathbf{x}^k) - H^* = O(1/k^2)$$

  accelerated version (FISTA) $\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{y}^k - t_k \nabla f(\mathbf{y}^k))$. Other multi-step methods exist (Tseng '10, Nesterov '13)

- (f strongly convex $C^{1,1}$)

$$H(\mathbf{x}^k) - H^* = O(q^k), q \in (0,1)$$

- (f nonconvex $C^{1,1}$) limit points are stationary points. $O(1/\sqrt{k})$ rate of optimality measure $G_s(\mathbf{x}^k) = \frac{1}{s}\|\mathbf{x}^k - \mathrm{prox}_{sg}(\mathbf{x}^k - s\nabla f(\mathbf{x}^k))\|$ to 0.

Rates of Convergence - Non-Block Conditional Gradient

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Conditional Gradient

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k), \mathbf{p}(\mathbf{x}^k) \in \underset{\mathbf{p}}{\operatorname{argmin}}\langle\{\nabla f(\mathbf{x}^k), \mathbf{p}\rangle + g(\mathbf{p})\}$$

g extended real-valued proper closed and convex

Rates of Convergence - Non-Block Conditional Gradient

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Conditional Gradient

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k), \mathbf{p}(\mathbf{x}^k) \in \underset{\mathbf{p}}{\operatorname{argmin}}\langle\{\nabla f(\mathbf{x}^k), \mathbf{p}\rangle + g(\mathbf{p})\}$$

g extended real-valued proper closed and convex

- Mostly useful when the prox is difficult to compute.
- $O(1/k)$ rate of convergence in the original Frank-Wolfe paper [56'] for g=indicator of polyhedral sets. [Levitin and Polyak '66] - extension to arbitrary compact convex sets. Extension to general g [Bach, 15']

Rates of Convergence - Non-Block Conditional Gradient

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Conditional Gradient

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k), \mathbf{p}(\mathbf{x}^k) \in \underset{\mathbf{p}}{\text{argmin}}\langle\{\nabla f(\mathbf{x}^k), \mathbf{p}\rangle + g(\mathbf{p})\}$$

g extended real-valued proper closed and convex

- Mostly useful when the prox is difficult to compute.
- $O(1/k)$ rate of convergence in the original Frank-Wolfe paper [56'] for g=indicator of polyhedral sets. [Levitin and Polyak '66] - extension to arbitrary compact convex sets. Extension to general g [Bach, 15']
- **Bad news**
  - Cannon and Culum ['68] – $O(1/k)$ is tight even for strongly convex functions.
  - Unknown if the method can be accelerated.

Rates of Convergence - Non-Block Conditional Gradient

$$H^* = \min\{H(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

Conditional Gradient

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k), \mathbf{p}(\mathbf{x}^k) \in \underset{\mathbf{p}}{\text{argmin}}\langle\{\nabla f(\mathbf{x}^k), \mathbf{p}\rangle + g(\mathbf{p})\}$$

g extended real-valued proper closed and convex

- Mostly useful when the prox is difficult to compute.
- $O(1/k)$ rate of convergence in the original Frank-Wolfe paper [56'] for g=indicator of polyhedral sets. [Levitin and Polyak '66] - extension to arbitrary compact convex sets. Extension to general g [Bach, 15']
- **Bad news**
  - Cannon and Culum ['68] – $O(1/k)$ is tight even for strongly convex functions.
  - Unknown if the method can be accelerated.
- $O(1/\sqrt{k})$ rate of the optimality measure $S(\mathbf{x}) = \langle\nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}(\mathbf{x})\rangle + g(\mathbf{x}) - g(\mathbf{p}(\mathbf{x}))$

Underlying Assumptions

- f is convex (most of the time...)
- ∇f is block-coordinate-wise Lipschitz continuous with local Lipschitz constants L_i:

$$\|\nabla_i f(\mathbf{x} + \mathbf{U}_i \mathbf{h}_i) - \nabla_i f(\mathbf{x})\| \leq L_i \|\mathbf{h}_i\|, \quad \text{for every } \mathbf{h}_i \in \mathbb{R}^{n_i}.$$

- (consequently) ∇f is Lipschitz continuous. Its constant is denoted by L.

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

- $S = \{\mathbf{x} : F(\mathbf{x}) \leq F(\mathbf{x}_0)\}$ is compact and we denote

$$R(\mathbf{x}_0) \equiv \max_{\mathbf{x} \in \mathbb{R}^n} \max_{\mathbf{x}^* \in X^*} \left\{\|\mathbf{x} - \mathbf{x}^*\| : F(\mathbf{x}) \leq F(\mathbf{x}_0)\right\}.$$

Summary of Rates of Convergence

	Cyclic		Randomized	
Method	NA	A	NA	A
Block PG	✓[1,6,7]	? [1]	✓[2,3]	✓[2,3]
Block CG	✓[4]	x	✓[5]	x

- **Cyclic** – the index i_k is chosen by the order $1, 2, \ldots, p, 1, 2, \ldots$. Also covers cyclic shuffle
- **Randomized** – the index i_k is chosen at random from $\{1, 2, \ldots, p\}$ at each iteration.
- **A** – accelerated $O(1/k^2)$ result. **NA** – non-accelerated $O(1/k)$ result.
- [1] = [B. Tetruashvili '13] ?= unconstrained setting [2] = [Fercoq, Richtarik '13], [3] = [Lin, Lu, Xiao '15] [4] = [B., Pauwels, Sabach '15], [5] = [Lacoste-Julien, Jaggi and Schmidt '13], [6] = [Shefi, Teboulle '16], [7]=[Hong, Wang, Razaviyayn, Luo '15]

Summary of Rates of Convergence

	Cyclic		Randomized	
Method	NA	A	NA	A
Block PG	✓[1,6,7]	? [1]	✓[2,3]	✓[2,3]
Block CG	✓[4]	x	✓[5]	x

- **Cyclic** – the index i_k is chosen by the order $1, 2, \ldots, p, 1, 2, \ldots$. Also covers cyclic shuffle
- **Randomized** – the index i_k is chosen at random from $\{1, 2, \ldots, p\}$ at each iteration.
- **A** – accelerated $O(1/k^2)$ result. **NA** – non-accelerated $O(1/k)$ result.
- [1] = [B. Tetruashvili '13] ?= unconstrained setting [2] = [Fercoq, Richtarik '13], [3] = [Lin, Lu, Xiao '15] [4] = [B., Pauwels, Sabach '15], [5] = [Lacoste-Julien, Jaggi and Schmidt '13], [6] = [Shefi, Teboulle '16], [7]=[Hong, Wang, Razaviyayn, Luo '15]
- Constants unfortunately depend on L or $\max\{L_1, L_2, \ldots, L_p\}$.

Summary of Rates of Convergence

	Cyclic		Randomized	
Method	NA	A	NA	A
Block PG	✓[1,6,7]	? [1]	✓[2,3]	✓[2,3]
Block CG	✓[4]	x	✓[5]	x

- **Cyclic** – the index i_k is chosen by the order $1, 2, \ldots, p, 1, 2, \ldots$. Also covers cyclic shuffle
- **Randomized** – the index i_k is chosen at random from $\{1, 2, \ldots, p\}$ at each iteration.
- **A** – accelerated $O(1/k^2)$ result. **NA** – non-accelerated $O(1/k)$ result.
- [1] = [B. Tetruashvili '13] ?= unconstrained setting [2] = [Fercoq, Richtarik '13], [3] = [Lin, Lu, Xiao '15] [4] = [B., Pauwels, Sabach '15], [5] = [Lacoste-Julien, Jaggi and Schmidt '13], [6] = [Shefi, Teboulle '16], [7]=[Hong, Wang, Razaviyayn, Luo '15]
- Constants unfortunately depend on L or $\max\{L_1, L_2, \ldots, L_p\}$.
- Possible to prove $O(1/\sqrt{k})$ rate of convergence of the optimality measures to 0 in the nonconvex case and $O(1/k)$ in the convex case.

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.
- Analysis of the randomized methods is usually much simpler. Sometimes even a simple adaptation of the non-block analysis.

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.
- Analysis of the randomized methods is usually much simpler. Sometimes even a simple adaptation of the non-block analysis.

Gradient Method
$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}^k)\|^2$

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.
- Analysis of the randomized methods is usually much simpler. Sometimes even a simple adaptation of the non-block analysis.

Gradient Method
$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}^k)\|^2$
- **B. Subgradient inequality+CS**
  $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{x}^*) \leq R\|f(\mathbf{x}^k)\|$

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.
- Analysis of the randomized methods is usually much simpler. Sometimes even a simple adaptation of the non-block analysis.

Gradient Method
$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}^k)\|^2$
- **B. Subgradient inequality+CS** $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{x}^*) \leq R\|f(\mathbf{x}^k)\|$
- A+B $\Rightarrow$ $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2LR^2}(f(\mathbf{x}^k) - f(\mathbf{x}^*))^2$

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.
- Analysis of the randomized methods is usually much simpler. Sometimes even a simple adaptation of the non-block analysis.

Gradient Method
$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}^k)\|^2$
- **B. Subgradient inequality+CS** $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{x}^*) \leq R\|f(\mathbf{x}^k)\|$
- A+B $\Rightarrow$ $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2LR^2}(f(\mathbf{x}^k) - f(\mathbf{x}^*))^2$
- Lemma: $a_k - a_{k+1} \geq \gamma a_k^2$ implies $a_k \leq \frac{1}{\gamma k}$
- $f(\mathbf{x}^k) - f^* \leq \frac{2LR^2}{k}$

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.
- Analysis of the randomized methods is usually much simpler. Sometimes even a simple adaptation of the non-block analysis.

Gradient Method
$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}^k)\|^2$
- **B. Subgradient inequality+CS** $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{x}^*) \leq R\|f(\mathbf{x}^k)\|$
- A+B $\Rightarrow$ $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2LR^2}(f(\mathbf{x}^k) - f(\mathbf{x}^*))^2$
- Lemma: $a_k - a_{k+1} \geq \gamma a_k^2$ implies $a_k \leq \frac{1}{\gamma k}$
- $f(\mathbf{x}^k) - f^* \leq \frac{2LR^2}{k}$

Randomized Block Gradient
$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla_{i_k} f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L_{i_k}}\|\nabla_{i_k} f(\mathbf{x}^k)\|^2 \geq \frac{1}{2L_{\max}}\|\nabla_{i_k} f(\mathbf{x}^k)\|^2$

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.
- Analysis of the randomized methods is usually much simpler. Sometimes even a simple adaptation of the non-block analysis.

Gradient Method
$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}^k)\|^2$
- **B. Subgradient inequality+CS** $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{x}^*) \leq R\|f(\mathbf{x}^k)\|$
- A+B $\Rightarrow$ $f(\mathbf{x}^k)-f(\mathbf{x}^{k+1}) \geq \frac{1}{2LR^2}(f(\mathbf{x}^k)-f(\mathbf{x}^*))^2$
- Lemma: $a_k - a_{k+1} \geq \gamma a_k^2$ implies $a_k \leq \frac{1}{\gamma k}$
- $f(\mathbf{x}^k) - f^* \leq \frac{2LR^2}{k}$

Randomized Block Gradient
$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla_{i_k} f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L_{i_k}}\|\nabla_{i_k} f(\mathbf{x}^k)\|^2 \geq \frac{1}{2L_{\max}}\|\nabla_{i_k} f(\mathbf{x}^k)\|^2$
- $\mathrm{E}(f(\mathbf{x}^k)) - \mathrm{E}(f(\mathbf{x}^{k+1})) \geq \frac{1}{2pL_{\max}}\|\nabla f(\mathbf{x}^k)\|^2$
- **B. The same**
- A+B $\Rightarrow$ $\mathrm{E}(f(\mathbf{x}^k)) - \mathrm{E}(f(\mathbf{x}^{k+1})) \geq \frac{1}{2pL_{\max}R^2}(\mathrm{E}(f(\mathbf{x}^k)) - f^*)^2$

Deterministic Vs. Randomized

- The constants in the deterministic efficiency estimates are worse than the randomized versions.
- Not consistent with the practical performance.
- Analysis of the randomized methods is usually much simpler. Sometimes even a simple adaptation of the non-block analysis.

Gradient Method

$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}^k)\|^2$
- **B. Subgradient inequality+CS** $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{x}^*) \leq R\|f(\mathbf{x}^k)\|$
- A+B $\Rightarrow$ $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2LR^2}(f(\mathbf{x}^k) - f(\mathbf{x}^*))^2$
- Lemma: $a_k - a_{k+1} \geq \gamma a_k^2$ implies $a_k \leq \frac{1}{\gamma k}$
- $f(\mathbf{x}^k) - f^* \leq \frac{2LR^2}{k}$

Randomized Block Gradient

$(\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla_{i_k} f(\mathbf{x}^k))$

- **A. Sufficient decrease**: $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L_{i_k}}\|\nabla_{i_k} f(\mathbf{x}^k)\|^2 \geq \frac{1}{2L_{\max}}\|\nabla_{i_k} f(\mathbf{x}^k)\|^2$
- $\mathrm{E}(f(\mathbf{x}^k)) - \mathrm{E}(f(\mathbf{x}^{k+1})) \geq \frac{1}{2pL_{\max}}\|\nabla f(\mathbf{x}^k)\|^2$
- **B. The same**
- A+B $\Rightarrow$ $\mathrm{E}(f(\mathbf{x}^k)) - \mathrm{E}(f(\mathbf{x}^{k+1})) \geq \frac{1}{2pL_{\max}R^2}(\mathrm{E}(f(\mathbf{x}^k)) - f^*)^2$
- $\mathrm{E}(f(\mathbf{x}^k)) - f^* \leq \frac{2pL_{\max}R^2}{k}$

Numerical Results On Synthetic Data

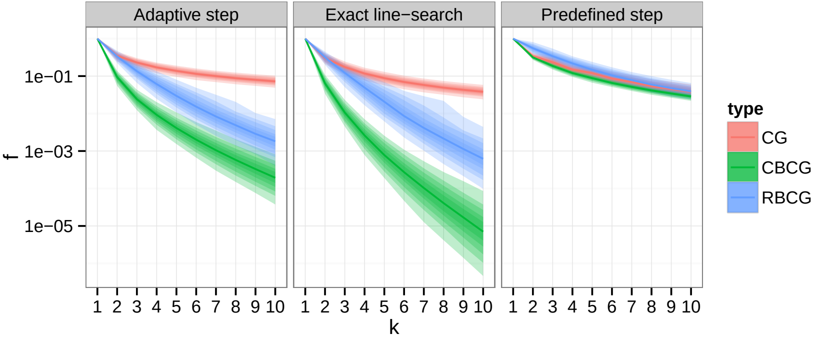
We solve the problem

$$\min_{\|\mathbf{x}\|_\infty \le 1} \frac{1}{2}(\mathbf{x}-\mathbf{y})^T\mathbf{Q}(\mathbf{x}-\mathbf{y}), \tag{1}$$

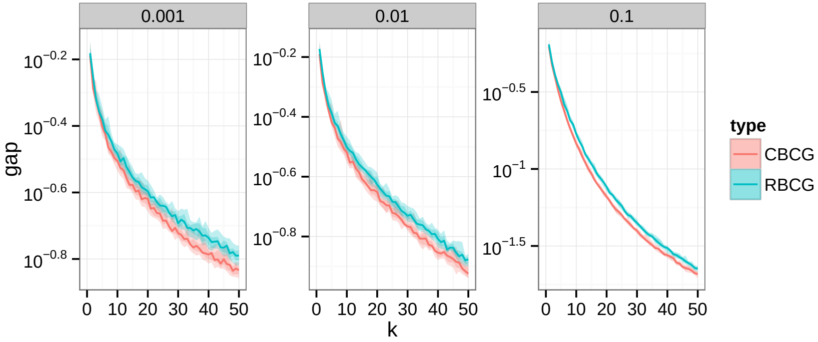
where

- $\mathbf{y} \in \mathbb{R}^{100}$ $\mathbf{Q} \in \mathbb{R}^{100\times 100}$.
- Generation of $\mathbf{Q}$: $\mathbf{Q} = \frac{1}{200}\mathbf{X}^T\mathbf{X}$ where each component of $\mathbf{X} \in \mathbb{R}^{200\times 100}$ is generated by $N(0,1)$.
- The entries of $\mathbf{y}$ are generated by $N(0,1)$.
- We compare the Conditional Gradient (CG), its random block version (RBDG) and its cyclic block version (CBCG) with the three different stepsize strategies based on 1000 randomly generated instances of problem.
- The central line is the median over the 1000 runs and the ribbons show 98%, 90%, 80%, 60% and 40% quantiles.
- k - number of effective passes through all the coordinates.

Comparison of Stepsize Rules and Methods



Results on Structural SVM



Results on the AM method with $p = 2$

- More difficult to analyze in the absence of strong convexity - distances between consecutive iterates cannot be controlled.
- On the other hand, logic dictates that if possible, exact minimization is better.
- Can theory substantiate this intuition?

Results on the AM method with $p = 2$

- More difficult to analyze in the absence of strong convexity - distances between consecutive iterates cannot be controlled.
- On the other hand, logic dictates that if possible, exact minimization is better.
- Can theory substantiate this intuition? Yes, at least for $p = 2$...

The General Minimization Problem

$$(\text{P}): \min \{H(\mathbf{y},\mathbf{z}) \equiv f(\mathbf{y},\mathbf{z}) + g_1(\mathbf{y}) + g_2(\mathbf{z}) : \mathbf{y} \in \mathbb{R}^{n_1}, \mathbf{z} \in \mathbb{R}^{n_2}\}$$

A. $g_1 : \mathbb{R}^{n_1} \to (-\infty,\infty], g_2 : \mathbb{R}^{n_2} \to (-\infty,\infty]$ are closed, proper and convex

B. f - convex and continuously differentiable function over $\text{dom}\, g_1 \times \text{dom}\, g_2$.

Block Notation

- $\mathbf{x} = (\mathbf{y}, \mathbf{z})$
- $g : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to (-\infty, \infty]$ is defined by
  $g(\mathbf{x}) = g(\mathbf{y}, \mathbf{z}) \equiv g_1(\mathbf{y}) + g_2(\mathbf{z})$.
- In this notation: $H(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$
- $\nabla_1 f(\mathbf{x})$ - gradient of f w.r.t $\mathbf{y}$ and $\nabla_2 f(\mathbf{x})$ - gradient w.r.t to $\mathbf{z}$.

Further Assumptions

C. $\nabla_1 f$ (uniformly) Lipschitz continuous w.r.t. to $\mathbf{y}$ over $\text{dom}\, g_1$ with constant $L_1 \in (0, \infty)$:

$$\|\nabla_1 f(\mathbf{y} + \mathbf{d}_1, \mathbf{z}) - \nabla_1 f(\mathbf{y}, \mathbf{z})\| \leq L_1 \|\mathbf{d}_1\|, \quad \mathbf{y}, \mathbf{y} + \mathbf{d}_1 \in \text{dom}\, g_1, \mathbf{z} \in \text{dom}\, g_2$$

D. $\nabla_2 f$ (uniformly) Lipschitz continuous w.r.t. to $\mathbf{z}$ over $\text{dom}\, g_2$ with constant $L_2 \in (0, \infty]$:

$$\|\nabla_2 f(\mathbf{y}, \mathbf{z} + \mathbf{d}_2) - \nabla_1 f(\mathbf{y}, \mathbf{z})\| \leq L_2 \|\mathbf{d}_2\|, \quad \mathbf{y} \in \text{dom}\, g_1, \mathbf{z}, \mathbf{z} + \mathbf{d}_2 \in \text{dom}\, g_2$$

When $L_2 = \infty$, [D] is meaningless!

The Alternating Minimization Method

Initialization: $\mathbf{y}_0 \in \text{dom}\, g_1, \mathbf{z}_0 \in \text{dom}\, g_2$ such that $\mathbf{z}_0 \in \underset{\mathbf{z}\in\mathbb{R}^{n_2}}{\text{argmin}}\, f(\mathbf{y}_0, \mathbf{z}) + g_2(\mathbf{z})$.

General Step (k=0,1,. . .):

$$\begin{aligned}
\mathbf{y}_{k+1} &\in \underset{\mathbf{y}\in\mathbb{R}^{n_1}}{\text{argmin}}\, f(\mathbf{y}, \mathbf{z}_k) + g_1(\mathbf{y}), \\
\mathbf{z}_{k+1} &\in \underset{\mathbf{z}\in\mathbb{R}^{n_2}}{\text{argmin}}\, f(\mathbf{y}_{k+1}, \mathbf{z}) + g_2(\mathbf{z}).
\end{aligned}$$

E. The optimal set of (P), denoted X^* is nonempty. The minimization problems

$$\min_{\mathbf{z}\in\mathbb{R}^{n_2}} f(\tilde{\mathbf{y}}, \mathbf{z}) + g_2(\mathbf{z}), \min_{\mathbf{y}\in\mathbb{R}^{n_1}} f(\mathbf{y}, \tilde{\mathbf{z}}) + g_1(\mathbf{y})$$

have minimizers for any $\tilde{\mathbf{y}} \in \text{dom}\, g_1, \tilde{\mathbf{z}} \in \text{dom}\, g_2$.

Note: A “half” step is performed before invoking the method.

Sublinear Rate of Convergence of AM

Theorm. For all $n \geq 2$

$$H(\mathbf{x}_n) - H^* \leq \max\left\{ \left(\frac{1}{2}\right)^{\frac{n-1}{2}} (H(\mathbf{x}_0) - H^*), \frac{8\min\{L_1, L_2\}R^2}{n-1} \right\}.$$

An ε-optimal solution is obtained after at most

$$\max\left\{ \frac{2}{\ln(2)}(\ln(H(\mathbf{x}_0) - H^*) + \ln(1/\varepsilon)), \frac{8\min\{L_1, L_2\}R^2}{\varepsilon} \right\} + 2$$

iterations.

- constant depends on $\min\{L_1, L_2\}$ - an optimistic result. The rate is dictated by the "best" function.
- weak dependence on global Lipschitz constants.

IRLS - Sublinear Rate of Convergence

$$(A) \quad \begin{array}{ll} \min & h_\eta(\mathbf{y},\mathbf{z}) \equiv s(\mathbf{y}) + \frac{1}{2}\sum_{i=1}^m \left(\frac{\|\mathbf{A}_i\mathbf{y}+\mathbf{b}_i\|^2+\eta^2}{z_i} + z_i \right) \\ \text{s.t.} & \mathbf{y} \in X \\ & \mathbf{z} \in [\eta/2,\infty)^m, \end{array} .$$

IRLS - Sublinear Rate of Convergence

$$(A) \quad \begin{array}{ll} \min & h_\eta(\mathbf{y},\mathbf{z}) \equiv s(\mathbf{y}) + \frac{1}{2}\sum_{i=1}^m \left(\frac{\|\mathbf{A}_i\mathbf{y}+\mathbf{b}_i\|^2+\eta^2}{z_i} + z_i \right) \\ \text{s.t.} & \mathbf{y} \in X \\ & \mathbf{z} \in [\eta/2, \infty)^m, \end{array} .$$

-

$$\begin{aligned} L_1 &= L_{\nabla s} + \frac{1}{\eta}\lambda_{\max}\left(\sum_{i=1}^m \mathbf{A}_i^T\mathbf{A}_i\right) \\ L_2 &= \infty \end{aligned}$$

$$(A) \quad \begin{array}{ll} \min & h_\eta(\mathbf{y},\mathbf{z}) \equiv s(\mathbf{y}) + \frac{1}{2}\sum_{i=1}^m \left(\frac{\|\mathbf{A}_i\mathbf{y}+\mathbf{b}_i\|^2+\eta^2}{z_i} + z_i \right) \\ \text{s.t.} & \mathbf{y} \in X \\ & \mathbf{z} \in [\eta/2, \infty)^m, \end{array} .$$

-

$$\begin{aligned} L_1 &= L_{\nabla s} + \frac{1}{\eta}\lambda_{\max}\left(\sum_{i=1}^m \mathbf{A}_i^T\mathbf{A}_i\right) \\ L_2 &= \infty \end{aligned}$$

- **Sublinear rate of convergence of IRLS:**

$$S_\eta(\mathbf{y}_n) - S_\eta^* \le \max\left\{ \left(\frac{1}{2}\right)^{\frac{n-1}{2}} (S_\eta(\mathbf{y}_0) - S_\eta^*), \frac{8L_1R^2}{n-1} \right\}.$$

Asymptotic Rate of Convergence

Theorem. There exists $K > 0$ such that

$$S_\eta(\mathbf{y}_n) - S_\eta^* \leq \frac{48R^2}{\eta(n-K)}$$

for all $n \geq K + 1$.

- Rate does not depend on the data (s,$\mathbf{A}_i, \mathbf{b}_i$).
- Possibly explains the fast empirical convergence of IRLS.

Duality?

Main Question: How does a dual-based variables decomposition method look like?

The Model

$$\min_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \sum_{i=1}^{p} \psi_i(\mathbf{x}) \right\},$$

- $f : \mathbb{E} \to (-\infty, \infty]$ is a closed, proper extended valued σ-strongly convex function.
- $\psi_i : \mathbb{E} \to (-\infty, \infty]$ $(i = 1, 2, \ldots, p)$ closed, proper extended real-valued convex.
- $\text{ri}(\text{dom}\, f) \cap (\cap_{i=1}^{p} \text{ri}(\text{dom}\, \psi_i)) \neq \emptyset$.

Functional Decomposition - the Idea

At each iteration of a functional decomposition method an operation involving only at most **one** of the functions ψ_i is performed.

Functional Decomposition - the Idea

At each iteration of a functional decomposition method an operation involving only at most **one** of the functions ψ_i is performed.

Suppose that either

- problems of the form $\min_{\mathbf{x}} f(\mathbf{x}) + \psi_i(\mathbf{x}) + \langle \mathbf{a}, \mathbf{x} \rangle$ can be easily solved.

or

- prox_{ψ_i} can be easily computed.

Functional Decomposition - the Idea

At each iteration of a functional decomposition method an operation involving only at most **one** of the functions ψ_i is performed.

Suppose that either

- problems of the form $\min_{\mathbf{x}} f(\mathbf{x}) + \psi_i(\mathbf{x}) + \langle \mathbf{a}, \mathbf{x} \rangle$ can be easily solved.

or

- prox_{ψ_i} can be easily computed.

Example of functional decomposition methods are incremental (sub)gradient methods (Kibradin [80'], Luo and Tseng[94'], Grippo [94'], Bertsekas [97'], Solodov [98'], Nedic and Bertsekas [00',01',10'], incremental subgradient-proximal (Bertsekas [10']) and certain variants of ADMM (Gabay and Mercier) and dual ADMM.

Example: 1D total variation denoising

Given a noisy measurements vector $\mathbf{y}$, we want to find a "smooth" vector $\mathbf{x}$ which is the solution to

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2 + \lambda \underbrace{\sum_{i=1}^{n-1} |x_i - x_{i+1}|}_{\psi(\mathbf{x})}.$$

- Equivalent to finding the prox of the TV function.
- ψ has no useful separability properties.

Example: 1D total variation denoising

Given a noisy measurements vector $\mathbf{y}$, we want to find a "smooth" vector $\mathbf{x}$ which is the solution to

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2 + \underbrace{\lambda \sum_{i=1}^{n-1} |x_i - x_{i+1}|}_{\psi(\mathbf{x})}.$$

- Equivalent to finding the prox of the TV function.
- ψ has no useful separability properties.

However, we can decompose ψ as $\psi = \psi_1 + \psi_2$ where

$$\begin{aligned} \psi_1(\mathbf{x}) &= \lambda \sum_{i=1}^{\lfloor n/2 \rfloor} |x_{2i-1} - x_{2i}| \\ \psi_2(\mathbf{x}) &= \lambda \sum_{i=1}^{\lfloor (n-1)/2 \rfloor} |x_{2i} - x_{2i+1}| \end{aligned}$$

$\mathrm{prox}_{\psi_1}, \mathrm{prox}_{\psi_2}$ can be easily computed since they are separable w.r.t. pair of variables (e.g., ψ_1 is separable w.r.t. to $\{x_1, x_2\}, \{x_3, x_4\}, \ldots$).

The Dual Problem

- The dual problem of (P) is

$$\text{(D)} \quad \max \left\{ q(\mathbf{y}) \equiv -f^* \left(-\sum_{j=1}^{p} \mathbf{y}_j \right) - \sum_{j=1}^{p} \psi_j^*(\mathbf{y}_j) \right\}$$

- In minimization form:

$$\min_{\mathbf{y} \in \mathbb{E}^p} \left\{ H(\mathbf{y}) \equiv F(\mathbf{y}) + \sum_{i=1}^{p} \Psi_i(\mathbf{y}_i) \right\}.$$

with $F(\mathbf{y}) \equiv f^* \left(-\sum_{j=1}^{p} \mathbf{y}_j \right)$ - a convex $C^{1,1}_{p/\sigma}$ function.
$\Psi_j(\mathbf{y}_j) \equiv \psi_j^*(\mathbf{y}_j)$ - closed, proper, convex.

The Dual Problem

- The dual problem of (P) is

$$\text{(D)} \quad \max \left\{ q(\mathbf{y}) \equiv -f^* \left(-\sum_{j=1}^p \mathbf{y}_j \right) - \sum_{j=1}^p \psi_j^*(\mathbf{y}_j) \right\}$$

- In minimization form:

$$\min_{\mathbf{y} \in \mathbb{E}^p} \left\{ H(\mathbf{y}) \equiv F(\mathbf{y}) + \sum_{i=1}^p \Psi_i(\mathbf{y}_i) \right\}.$$

with $F(\mathbf{y}) \equiv f^* \left(-\sum_{j=1}^p \mathbf{y}_j \right)$ - a convex $C^{1,1}_{p/\sigma}$ function.
$\Psi_j(\mathbf{y}_j) \equiv \psi_j^*(\mathbf{y}_j)$ - closed, proper, convex.

Dual block variables decomposition = primal functional decomposition

Two Dual Block Steps

Given $\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_p$, the objective is to compute $\bar{\mathbf{y}}_i^{\text{new}}$ - a new value of the ith component by employing one of the following steps:

- **dual exact minimization step.**

$$\mathbf{y}_i^{\text{new}} \in \operatorname{argmin}\left\{ f^*\left(-\sum_{j=1, j\neq i}^p \bar{\mathbf{y}}_j - \mathbf{y}_i\right) + \psi_i^*(\mathbf{y}_i)\right\}.$$

  - the value of $\bar{\mathbf{y}}_i$ is not being used.

- **dual proximal gradient step.**

$$\mathbf{y}_i^{\text{new}} = \mathrm{prox}_{\sigma\psi_i^*}(\bar{\mathbf{y}}_i + \sigma\nabla f^*(-\textstyle\sum_{j=1}^p \bar{\mathbf{y}}_j)).$$

Primal Representations of the Dual Block Steps:

Using Moreau decomposition and some conjugate/prox calculus...

Primal Representation of the Dual Exact Minimization Step:

$$\begin{aligned}\tilde{\mathbf{y}}_i &= \textstyle\sum_{j\neq i} \bar{\mathbf{y}}_j, \\ \bar{\mathbf{x}} &\in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \psi_i(\mathbf{x}) + \langle \tilde{\mathbf{y}}_i, \mathbf{x}\rangle \right\}, \\ \mathbf{y}_i^{\text{new}} &\in \partial\psi_i(\bar{\mathbf{x}}).\end{aligned}$$

Primal Representation of the Dual Proximal Gradient Step:

$$\begin{aligned}\bar{\mathbf{x}} &= \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \langle \textstyle\sum_{j=1}^{p} \bar{\mathbf{y}}_j, \mathbf{x}\rangle \right\}, \\ \bar{\mathbf{y}}_i^{\text{new}} &= \bar{\mathbf{y}}_i + \sigma\bar{\mathbf{x}} - \operatorname{prox}_{\psi_i/\sigma}\left(\frac{\bar{\mathbf{y}}_i}{\sigma} + \bar{\mathbf{x}}\right)\end{aligned}$$

Dual Cyclic Alternating Minimization Method (DAM-C)

Initialization. $\mathbf{y}^0 = (\mathbf{y}_0^0, \mathbf{y}_1^0, \ldots, \mathbf{y}_m^0) \in \mathbb{E}^p$.
General Step $(k = 0, 1, 2, 3, \ldots)$**.**

- Set $\mathbf{y}^{k,0} = \mathbf{y}^k$.
- For $i = 0, 1, \ldots, p-1$
  Define $\mathbf{y}^{k,i+1}$ as follows:

$$\mathbf{x}^{k,i} \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \psi_{i+1}(\mathbf{x}) + \langle \textstyle\sum_{j=1,j\neq i+1}^{p} \mathbf{y}_j^{k,i}, \mathbf{x}\rangle \right\}$$

$$\mathbf{y}_j^{k,i+1} \begin{cases} \in \partial\psi_{i+1}(\mathbf{x}^{k,i}) & j = i+1, \\ = \mathbf{y}_j^{k,i} & j \neq i+1. \end{cases}$$

- Set $\mathbf{y}^{k+1} = \mathbf{y}^{k,p}$ and $\mathbf{x}^k = \mathbf{x}^{k,0}$.

If $f \in C^1$, then the update rule for $\mathbf{y}^{k,i+1}$ can be replaced by

$$\mathbf{y}_j^{k,i+1} = \begin{cases} -\nabla f(\mathbf{x}^{k,i}) - \sum_{j=1,j\neq i+1}^{p} \mathbf{y}_j^{k,i} & j = i+1, \\ \mathbf{y}_j^{k,i} & j \neq i+1. \end{cases}$$

Dual Cyclic Block Proximal Gradient Method (DBPG-C)

Initialization. $(\mathbf{y}_0^0, \mathbf{y}_1^0, \ldots, \mathbf{y}_m^0) \in \mathbb{E}^p$.
General Step $(k = 0, 1, 2, 3, \ldots)$**.**

- Set $\mathbf{y}^{k,0} = \mathbf{y}^k$.
- For $i = 0, 1, \ldots, m-1$
  Define $\mathbf{y}^{k,i+1}$ as follows

$$\mathbf{x}^{k,i} = \underset{\mathbf{x}\in\mathbb{E}}{\operatorname{argmin}} \left\{ f(\mathbf{x}) + \langle \textstyle\sum_{j=1}^{p} \mathbf{y}_j^{k,i}, \mathbf{x} \rangle \right\},$$

$$\mathbf{y}_j^{k,i+1} = \begin{cases} \mathbf{y}_{i+1}^k + \sigma \mathbf{x}^{k,i} - \operatorname{prox}_{\psi_{i+1}/\sigma}\left(\frac{\mathbf{y}_{i+1}^{k,i}}{\sigma} + \mathbf{x}^{k,i} \right) & j = i+1, \\ \mathbf{y}_j^{k,i}, & j \neq i+1. \end{cases}$$

- Set $\mathbf{y}^{k+1} = \mathbf{y}^{k,m}$ and $\mathbf{x}^k = \mathbf{x}^{k,0}$.

Rate of Convergence of the Primal Sequence

- The rates of convergence of the dual objective function are already known.
- Does it imply corresponding rates of convergence of the primal sequence?

Rate of Convergence of the Primal Sequence

- The rates of convergence of the dual objective function are already known.
- Does it imply corresponding rates of convergence of the primal sequence? YES!

Rate of Convergence of the Primal Sequence

- The rates of convergence of the dual objective function are already known.
- Does it imply corresponding rates of convergence of the primal sequence? YES!

The primal-dual relation. Let $\bar{\mathbf{y}}$ satisfy $\bar{\mathbf{y}}_j \in \operatorname{dom} \psi_j^*$ for any $j \in \{1, 2, \ldots, p\}$. Let $\bar{\mathbf{x}}$ be defined by either

$$\bar{\mathbf{x}} \in \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ f(\mathbf{x}) + \langle \textstyle\sum_{i=j}^p \bar{\mathbf{y}}_j, \mathbf{x} \rangle \right\}$$

or

$$\bar{\mathbf{x}} \in \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ f(\mathbf{x}) + \psi_i(\mathbf{x}) + \langle \textstyle\sum_{j=1, j \neq i}^p \bar{\mathbf{y}}_j, \mathbf{x} \rangle \right\}$$

for some $i \in \{1, 2, \ldots, p\}$. Then

$$\|\bar{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\rm opt} - q(\bar{\mathbf{y}}))$$

Rates of Convergence of Functional Decomposition Methods

method	complexity result	remarks
DBPG-C	$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \le \frac{2C_1}{\sigma(k+1)}$	Ψ_i^* indicators, general m
DBPG-C	$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \le \frac{2C_2}{\sigma(k+1)}$	general Ψ_i and m
DBPG-R	$\mathbb{E}(\|\mathbf{x}^k - \mathbf{x}^*\|^2) \le \frac{2m}{\sigma(m+k)} C_3$	general Ψ_i and m
DAM-C	$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \le \frac{2C_4}{\sigma k}$	$m = 2$
DAM-C	$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \le \frac{2C_5}{\sigma(k+1)}$	general m and ψ_i

$$
\begin{aligned}
C_1 &= \frac{2m\left[(2m+1)R + \sigma M\right]^2}{\sigma} \\
C_2 &= 2\sigma m G_{\max}^2 R^2 \max\left\{\frac{2}{\sigma m G_{\max}^2 R^2} - 2, q_{\rm opt} - q(\mathbf{y}^0), 2\right\} \\
C_3 &= \frac{1}{2\sigma} \min_{\mathbf{y}^* \in Y^*} \|\mathbf{y}_0 - \mathbf{y}^*\|^2 + q_{\rm opt} - q(\mathbf{y}^0), \\
C_4 &= 3\max\left\{q_{\rm opt} - q(\mathbf{y}^0), \frac{1}{\sigma} R^2\right\}, \\
C_5 &= \frac{2m^3 R^2 \max\left\{\frac{2\sigma}{m^3 R^2} - 2, q_{\rm opt} - q(\mathbf{y}^0), 2\right\}}{\sigma}
\end{aligned}
$$

Numerical Example: Isotropic 2D TV denoising

- TV denoising:

$$\min_{\mathbf{x}\in\mathbb{R}^{m\times n}} \frac{1}{2}\|\mathbf{x}-\mathbf{b}\|_F^2 + \theta \cdot \mathrm{TV}_I(\mathbf{x})$$

- Isotropic TV:

$$\mathbf{x} \in \mathbb{R}^{m\times n} \quad \mathrm{TV}_I = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\sqrt{(x_{i,j}-x_{i+1,j})^2+(x_{i,j}-x_{i,j+1})^2} + \sum_{i=1}^{m-1}|x_{i,n}-x_{i+1,n}| + \sum_{j=1}^{n-1}|x_{m,j}-x_{m,j+1}|,$$

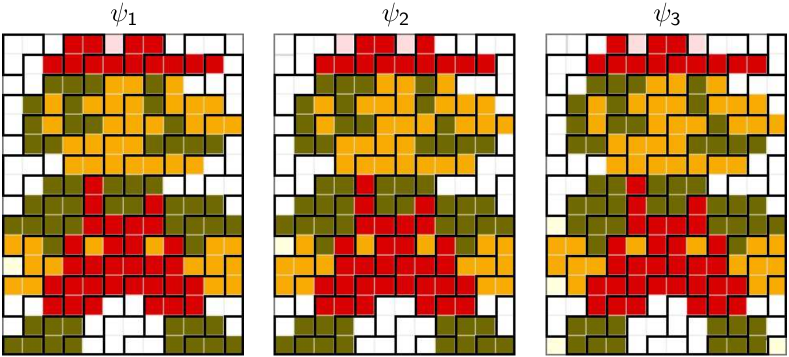
Chambolle, Pock[15'] : anisotropic ($l_1 - l_1$), decomposition intro rows and columns (two functions).

$$\begin{aligned}
\mathrm{TV}_I(\mathbf{x}) = & \quad \sum_{i=1}^m \sum_{j=1}^n \sqrt{(x_{i,j}-x_{i+1,j})^2+(x_{i,j}-x_{i,j+1})^2} \\
= & \quad \sum_{k\in K_1} \sum_{(i,j)\in D_k} \sqrt{(x_{i,j}-x_{i+1,j})^2+(x_{i,j}-x_{i,j+1})^2} \\
& + \sum_{k\in K_2} \sum_{(i,j)\in D_k} \sqrt{(x_{i,j}-x_{i+1,j})^2+(x_{i,j}-x_{i,j+1})^2} \\
& + \sum_{k\in K_3} \sum_{(i,j)\in D_k} \sqrt{(x_{i,j}-x_{i+1,j})^2+(x_{i,j}-x_{i,j+1})^2} \\
= & \quad \psi_1(\mathbf{x})+\psi_2(\mathbf{x})+\psi_3(\mathbf{x}).
\end{aligned}$$

D_k - indices of the k diagonal.

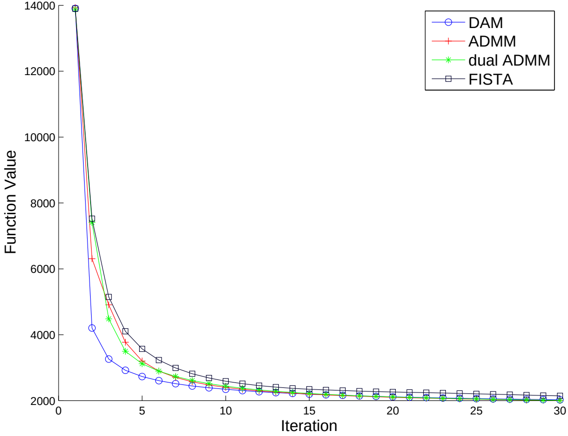
$$K_i \equiv \left\{k \in \{-(m-1),\ldots,n-1\} : (k+1-i) \mod 3 = 0\right\} \qquad i=1,2,3.$$

Using the separability of ψ_i, computation of prox_{ψ_i} is simple.


ψ_1
ψ_2
ψ_3

Numerical Comparison

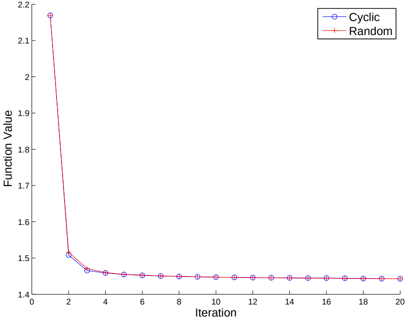
30 iterations, $\lambda = 0.5$.



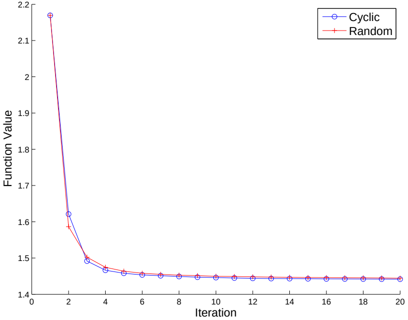
In the first 100 iterations DAM-C is better than FISTA. However, after "enough" runs, FISTA wins...

Random Versus Deterministic

$p = 3$



$p = 30$



Almost the same performance, with a slight advantage to the cyclic rule.

References

- Amir Beck and Luba Tetruashvili, "On The Convergence of Block Coordinate Descent Type Methods", *SIAM J. Optim.*, **23**, no. 4 (2013) 2037–2060.
- Amir Beck, "On the Convergence of Alternating Minimization for Convex Programming with Applications to Iteratively Reweighted Least Squares and Decomposition Schemes", *SIAM J. Optim.*, **25**, no. 1 (2015), 185–209.
- Amir Beck, Edouard Pauwels and Shoham Sabach, "The Cyclic Block Conditional Gradient Method for Convex Optimization Problems" (2015), submitted for publication.
- Amir Beck, Luba Tetruashvili, Yakov Vaisbourd, "Rate of Convergence Analysis of Dual-Based Variables Decomposition Method"

Last Slide

THANK YOU FOR YOUR ATTENTION