Primal and Dual
Predicted Decrease Approximation Methods

Amir Beck

Technion - Israel Institute of Technology
Haifa, Israel

Joint work with
Edouard Pauwels and Shoham Sabach

Workshop on Large-Scale and Distributed Optimization
Lund, Sweden, 14-16 June 2017

The Primal Model

$$\text{(Q)} \quad \min_{\mathbf{y}\in\mathbb{R}^d} \{H(\mathbf{y}) \equiv F(\mathbf{y}) + G(\mathbf{y})\}$$

- $F : \mathbb{R}^d \to \mathbb{R}$ is convex L-smooth over $\mathbb{R}^d$.
- $G : \mathbb{R}^d \to (-\infty, \infty]$ proper, closed, convex with a compact domain.

Two widely-used methods for solving (Q):

- Proximal gradient.
- Generalized conditional gradient.

Proximal Gradient

$$\mathbf{x}^{k+1} = \mathrm{prox}_{t_k G}\left(\mathbf{x}^k - t_k \nabla F(\mathbf{x}^k)\right).$$

- $O(1/k)$ rate of convergence in function values.
- faster rates of $O(1/k^2)$ are possible (e.g., FISTA [B. Teboulle 09'], accelerated methods [Nesterov, 13']). Under strong convexity even faster - $O(q^K)$.

Generalized Conditional Gradient

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k),$$

where

$$\mathbf{p}(\mathbf{x}^k) \in \underset{\mathbf{p}}{\operatorname{argmin}} \left\{ \langle \nabla F(\mathbf{x}^k), \mathbf{p} \rangle + G(\mathbf{p}) \right\}.$$

Idea: linearize F, keep G. Go towards the direction of the obtained vector $\mathbf{p}(\mathbf{x}^k)$.

Generalized Conditional Gradient

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k),$$

where

$$\mathbf{p}(\mathbf{x}^k) \in \operatorname*{argmin}_{\mathbf{p}} \left\{ \langle \nabla F(\mathbf{x}^k), \mathbf{p} \rangle + G(\mathbf{p}) \right\}.$$

Idea: linearize F, keep G. Go towards the direction of the obtained vector $\mathbf{p}(\mathbf{x}^k)$.

- If $G = \delta_C$, GCG amounts to the conditional gradient/Frank-Wolfe [56']. GCG was introduced in [Bach 15']
- $O(1/k)$ rate of convergence in function values.
- No acceleration if objective if strongly convex. [Canon, Cullum 68']

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k),$$

where

$$\mathbf{p}(\mathbf{x}^k) \in \underset{\mathbf{p}}{\operatorname{argmin}} \left\{ \langle \nabla F(\mathbf{x}^k), \mathbf{p} \rangle + G(\mathbf{p}) \right\}.$$

Idea: linearize F, keep G. Go towards the direction of the obtained vector $\mathbf{p}(\mathbf{x}^k)$.

- If $G = \delta_C$, GCG amounts to the conditional gradient/Frank-Wolfe [56']. GCG was introduced in [Bach 15']
- $O(1/k)$ rate of convergence in function values.
- No acceleration if objective if strongly convex. [Canon, Cullum 68']
- Analysis depends on the optimality measure

$$S(\mathbf{y}) \equiv \max_{\mathbf{p}} \{ \langle \nabla F(\mathbf{y}), \mathbf{y} - \mathbf{p} \rangle + G(\mathbf{y}) - G(\mathbf{p}) \}$$

Generalized Conditional Gradient

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k),$$

where

$$\mathbf{p}(\mathbf{x}^k) \in \operatorname*{argmin}_{\mathbf{p}} \left\{ \langle \nabla F(\mathbf{x}^k), \mathbf{p} \rangle + G(\mathbf{p}) \right\}.$$

Idea: linearize F, keep G. Go towards the direction of the obtained vector $\mathbf{p}(\mathbf{x}^k)$.

- If $G = \delta_C$, GCG amounts to the conditional gradient/Frank-Wolfe [56']. GCG was introduced in [Bach 15']
- $O(1/k)$ rate of convergence in function values.
- No acceleration if objective if strongly convex. [Canon, Cullum 68']
- Analysis depends on the optimality measure

$$S(\mathbf{y}) \equiv \max_{\mathbf{p}} \{ \langle \nabla F(\mathbf{y}), \mathbf{y} - \mathbf{p} \rangle + G(\mathbf{y}) - G(\mathbf{p}) \}$$

The Optimality Measure

$$S(\mathbf{y}) \equiv \max_{\mathbf{p}} \left\{ \langle \nabla F(\mathbf{y}), \mathbf{y} - \mathbf{p} \rangle + G(\mathbf{y}) - G(\mathbf{p}) \right\}.$$

Notation: $\mathbf{p}(\mathbf{y}) \in \operatorname*{argmin}_{\mathbf{p}} \left\{ \langle \nabla F(\mathbf{y}), \mathbf{p} \rangle + G(\mathbf{p}) \right\}$ (CG step)

Properties:

- $S(\mathbf{y}) \geq 0 \ \forall \mathbf{y}$ and $S(\mathbf{y}) = 0$ iff $\mathbf{y}$ is optimal.
- $H(\mathbf{y}) - H^* \leq S(\mathbf{y})$.
- $S(\mathbf{y}) = \langle \nabla F(\mathbf{y}), \mathbf{y} \rangle + G(\mathbf{y}) - [\langle \nabla F(\mathbf{y}), \mathbf{p}(\mathbf{y}) \rangle + G(\mathbf{p}(\mathbf{y}))]$.
  predicted decrease at $\mathbf{y}$ by the linearized function
  $\mathbf{z} \mapsto \langle \nabla F(\mathbf{y}), \mathbf{z} \rangle + G(\mathbf{z})$.

predicted decrease approximation

Definition. For $\gamma \geq 1$ and $\bar{\mathbf{y}} \in \text{dom}\, G$, a vector $\mathbf{u}(\bar{\mathbf{y}}) \in \text{dom}\, G$ is a $\frac{1}{\gamma}$-predicted decrease approximation (PDA) vector of H at $\bar{\mathbf{y}}$ if

$$\frac{1}{\gamma} S(\bar{\mathbf{y}}) \leq \langle \nabla F(\bar{\mathbf{y}}), \bar{\mathbf{y}} - \mathbf{u}(\bar{\mathbf{y}}) \rangle + G(\bar{\mathbf{y}}) - G(\mathbf{u}(\bar{\mathbf{y}})).$$

- $\frac{1}{\gamma}$ - approximation factor
- $\mathbf{u}(\bar{\mathbf{y}})$ captures at least a proportion of $S(\bar{\mathbf{y}})$.
- $\mathbf{u}(\bar{\mathbf{y}}) = \mathbf{p}(\bar{\mathbf{y}})$ - 1-PDA vector.
- Simple generalization of the notion of "approximate linear oracle" with multiplicative error [Lacoste-Julien et al 13'].

predicted decrease approximation

Definition. For $\gamma \geq 1$ and $\bar{\mathbf{y}} \in \mathrm{dom}\, G$, a vector $\mathbf{u}(\bar{\mathbf{y}}) \in \mathrm{dom}\, G$ is a $\frac{1}{\gamma}$-predicted decrease approximation (PDA) vector of H at $\bar{\mathbf{y}}$ if

$$\frac{1}{\gamma} S(\bar{\mathbf{y}}) \leq \langle \nabla F(\bar{\mathbf{y}}), \bar{\mathbf{y}} - \mathbf{u}(\bar{\mathbf{y}}) \rangle + G(\bar{\mathbf{y}}) - G(\mathbf{u}(\bar{\mathbf{y}})).$$

- $\frac{1}{\gamma}$ - approximation factor
- $\mathbf{u}(\bar{\mathbf{y}})$ captures at least a proportion of $S(\bar{\mathbf{y}})$.
- $\mathbf{u}(\bar{\mathbf{y}}) = \mathbf{p}(\bar{\mathbf{y}})$ - 1-PDA vector.
- Simple generalization of the notion of "approximate linear oracle" with multiplicative error [Lacoste-Julien et al 13'].
- The point is **not** that actual errors occur in the oracle evaluation, but the notion **allows to ensure additional structure in the form of the update while maintaining desirable convergence properties.**

Example: Block Separable G, 1-Sparse Updates

Setting:

- partition: $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m), \mathbf{y}_i \in \mathbb{R}^{d_i}$.
- $m = d_1 + d_2 + \dots + d_m$.
- $G(\mathbf{y}) = \sum_{i=1}^m G_i(\mathbf{y}_i)$.

Main observation in this example: Given $\bar{\mathbf{y}} \in \mathbb{R}^m$, it is possible to find a $\frac{1}{m}$-PDA vector different than $\bar{\mathbf{y}}$ in only one component.

Example: Block Separable G, 1-Sparse Updates

Setting:

- partition: $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m), \mathbf{y}_i \in \mathbb{R}^{d_i}$.
- $m = d_1 + d_2 + \ldots + d_m$.
- $G(\mathbf{y}) = \sum_{i=1}^m G_i(\mathbf{y}_i)$.

Main observation in this example: Given $\bar{\mathbf{y}} \in \mathbb{R}^m$, it is possible to find a $\frac{1}{m}$-PDA vector different than $\bar{\mathbf{y}}$ in only one component.

- **partial optimality measures:**

$$\begin{aligned} S_i(\mathbf{y}) &= \max_{\mathbf{p}_i} \left\{ \langle \nabla_i F(\mathbf{y}), \mathbf{y}_i - \mathbf{p}_i \rangle + G_i(\mathbf{y}_i) - G_i(\mathbf{p}_i) \right\}, \\ \mathbf{p}_i(\mathbf{y}) &\in \operatorname*{argmin}_{\mathbf{p}_i} \left\{ \langle \nabla_i F(\mathbf{y}), \mathbf{p}_i \rangle + G_i(\mathbf{p}_i) \right\}. \end{aligned}$$

Computation of a 1-sparse $\frac{1}{m}$-PDA vector:

- **Define** $\bar{i} \in \operatorname*{argmax}_{i=1,2,\ldots,m} S_i(\bar{\mathbf{y}})$.
- $\mathbf{u}(\bar{\mathbf{y}})_j = \bar{\mathbf{y}}_j (j \neq \bar{i})$, $\mathbf{u}(\bar{\mathbf{y}})_{\bar{i}} = \mathbf{p}_{\bar{i}}(\bar{\mathbf{y}})$.

The $\frac{1}{\gamma}$-PDA Method

Initialization. $\mathbf{y}^0 \in \text{dom } G$.
General Step. For $k = 0, 1, \ldots,$

(i)
- Choose $\mathbf{u}(\mathbf{y}^k)$- a $\frac{1}{\gamma}$-PDA vector of H at $\mathbf{y}^k$.
- Choose compact X^k s.t. $[\mathbf{y}^k, \mathbf{u}(\mathbf{y}^k)] \subseteq X^k$.

(ii) Perform one of the following:

$$\textbf{prox-grad update: } \mathbf{y}^{k+1} = \mathrm{prox}_{\frac{1}{L_k} G + \delta_{X^k}} \left(\mathbf{y}^k - \frac{1}{L_k} \nabla F(\mathbf{y}^k) \right) \tag{1}$$

$$\textbf{exact update: } \mathbf{y}^{k+1} = \underset{\mathbf{y} \in X^k}{\text{argmin}}\, F(\mathbf{y}) + G(\mathbf{y}) \tag{2}$$

- L_k is chosen to satisfy

$$F(\mathbf{y}^{k+1}) \leq F(\mathbf{y}^k) + \langle \nabla F(\mathbf{y}^k), \mathbf{y}^{k+1} - \mathbf{y}^k \rangle + \frac{L_k}{2} \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2.$$

Example: Generalized Conditional Gradient [Bach, 15']

Initialization: $\mathbf{y}^0 \in \operatorname{dom} G$.
General step (k=0,1,...):

- Compute $\mathbf{p}(\mathbf{y}^k) \in \underset{\mathbf{p}}{\operatorname{argmin}} \left\{ \langle \nabla f(\mathbf{y}^k), \mathbf{p} \rangle + G(\mathbf{p}) \right\}$.
- Set $\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)$ where

$$t_k \in \underset{t \in [0,1]}{\operatorname{argmin}} H(\mathbf{y}^k + t(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)).$$

- 1-PDA method – $\mathbf{u}(\mathbf{y}) = \mathbf{p}(\mathbf{y}), X^k = [\mathbf{y}^k, \mathbf{u}(\mathbf{y}^k)]$.

Example: Generalized Conditional Gradient [Bach, 15']

Initialization: $\mathbf{y}^0 \in \mathrm{dom}\, G$.
General step (k=0,1,. . .):

- Compute $\mathbf{p}(\mathbf{y}^k) \in \underset{\mathbf{p}}{\operatorname{argmin}} \left\{ \langle \nabla f(\mathbf{y}^k), \mathbf{p} \rangle + G(\mathbf{p}) \right\}$.
- Set $\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)$ where

$$t_k \in \underset{t \in [0,1]}{\operatorname{argmin}} H(\mathbf{y}^k + t(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)).$$

- 1-PDA method – $\mathbf{u}(\mathbf{y}) = \mathbf{p}(\mathbf{y}), X^k = [\mathbf{y}^k, \mathbf{u}(\mathbf{y}^k)]$.
- Changing X^k to $X^k = \{\mathbf{y}^k + t(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k) : t \geq 0\}$, we can take larger stepsizes:

$$t_k \in \underset{t \in [0,\infty)}{\operatorname{argmin}} H(\mathbf{y}^k + t(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k)).$$

Example 2: Proximal Gradient

Initialization: $\mathbf{y}^0 \in \text{dom}\, G$.
General step (k=0,1,...):

- compute

$$\mathbf{y}^{k+1} = \text{prox}_{\frac{1}{L_k} G} \left(\mathbf{y}^k - \frac{1}{L_k} \nabla F(\mathbf{y}^k) \right).$$

- 1-PDA method.
- $\mathbf{u}(\mathbf{y}) \equiv \mathbf{p}(\mathbf{y}), X^k = \mathbb{R}^d$
- constant stepsize, backtracking
- possible extension to hybrid proximal gradient/generalized conditional gradient

Example 3: Greedy CD for separable problems

Setting: $G(\mathbf{y}) = \sum_{i=1}^{m} G_i(\mathbf{y}_i)$.
Two possible choices for X^k (both contain a $1/m$-PDA vector)

$$\begin{aligned}
\bar{X}^k &= \{\bar{\mathbf{y}}_1\} \times \{\bar{\mathbf{y}}_2\} \times \cdots \{\bar{\mathbf{y}}_{\bar{i}-1}\} \times [\mathbf{y}_{\bar{i}}^k, \mathbf{p}_{\bar{i}}(\mathbf{y}^k)] \times \{\bar{\mathbf{y}}_{\bar{i}+1}\} \times \cdots \times \{\bar{\mathbf{y}}_m\}, \\
\tilde{X}^k &= \{\bar{\mathbf{y}}_1\} \times \{\bar{\mathbf{y}}_2\} \times \cdots \{\bar{\mathbf{y}}_{\bar{i}-1}\} \times \operatorname{dom} G_{\bar{i}} \times \{\bar{\mathbf{y}}_{\bar{i}+1}\} \times \cdots \times \{\bar{\mathbf{y}}_m\}.
\end{aligned}$$

Example 3: Greedy CD for separable problems

Setting: $G(\mathbf{y}) = \sum_{i=1}^m G_i(\mathbf{y}_i)$.
Two possible choices for X^k (both contain a $1/m$-PDA vector)

$$\begin{aligned}\bar{X}^k &= \{\bar{\mathbf{y}}_1\} \times \{\bar{\mathbf{y}}_2\} \times \cdots \{\bar{\mathbf{y}}_{\bar{i}-1}\} \times [\mathbf{y}_{\bar{i}}^k, \mathbf{p}_{\bar{i}}(\mathbf{y}^k)] \times \{\bar{\mathbf{y}}_{\bar{i}+1}\} \times \cdots \times \{\bar{\mathbf{y}}_m\},\\ \tilde{X}^k &= \{\bar{\mathbf{y}}_1\} \times \{\bar{\mathbf{y}}_2\} \times \cdots \{\bar{\mathbf{y}}_{\bar{i}-1}\} \times \operatorname{dom} G_{\bar{i}} \times \{\bar{\mathbf{y}}_{\bar{i}+1}\} \times \cdots \times \{\bar{\mathbf{y}}_m\}.\end{aligned}$$

Initialization: $\mathbf{y}^0 \in \operatorname{dom} G$.
General step $(k = 0, 1, \ldots)$**:**

- Compute $\bar{i} \in \underset{i=1,2,\ldots,m}{\operatorname{argmax}} S_i(\mathbf{y}^k)$, where

$$S_i(\mathbf{y}^k) = \langle \nabla F_i(\mathbf{y}^k), \mathbf{y}_i^k - \mathbf{p}_i(\mathbf{y}^k)\rangle + G_i(\mathbf{y}_i^k) - G_i(\mathbf{p}_i(\mathbf{y}^k))$$

  with $\mathbf{p}_i(\mathbf{y}^k) \in \underset{\mathbf{p}_i}{\operatorname{argmin}}\{\langle \nabla F_i(\mathbf{y}^k), \mathbf{p}_i\rangle + G_i(\mathbf{p}_i)\}$.

- **Core step:** Compute $\mathbf{y}^{k+1}$.

Example 3: Greedy coordinate descent for separable problems

The update formula of $\mathbf{y}^{k+1}$ ("core step") depends on X^k and the type of update rule (exact/prox-grad)

Example 3: Greedy coordinate descent for separable problems

The update formula of $\mathbf{y}^{k+1}$ ("core step") depends on X^k and the type of update rule (exact/prox-grad)

- **greedy block CG** ($X^k = \bar{X}^k$, exact update)

$$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k \mathbf{U}_{\bar{i}}(\mathbf{p}_{\bar{i}}(\mathbf{y}^k) - \mathbf{y}^k_{\bar{i}}),$$

where $t_k \in \operatorname*{argmin}_{0\le t\le 1} H\left(\mathbf{y}^k + t\mathbf{U}_{\bar{i}}(\mathbf{p}_{\bar{i}}(\mathbf{y}^k) - \mathbf{y}^k_{\bar{i}})\right)$.

- **greedy block minimization** ($X^k = \tilde{X}^k$, exact update)

$$\mathbf{y}_i^{k+1} \begin{cases} = \mathbf{y}_i^k, & i \neq \bar{i}, \\ \in \operatorname*{argmin}_{\mathbf{y}_{\bar{i}}} \left\{ F(\mathbf{y}^k + \mathbf{U}_{\bar{i}}(\mathbf{y}_{\bar{i}} - \mathbf{y}^k_{\bar{i}})) + G_{\bar{i}}(\mathbf{y}_{\bar{i}}) : \mathbf{y}_{\bar{i}} \in \operatorname{dom} G_{\bar{i}} \right\}, & i = \bar{i}. \end{cases}$$

- **greedy block proximal-gradient** ($X^k = \tilde{X}^k$, prox-grad step)

$$\mathbf{y}_i^{k+1} = \begin{cases} \mathbf{y}_i^k, & i \neq \bar{i}, \\ \operatorname{prox}_{\frac{1}{L_k} G_{\bar{i}}}\left(\mathbf{y}^k_{\bar{i}} - \frac{1}{L_k}\nabla_{\bar{i}} F(\mathbf{y}^k)\right), & i = \bar{i}. \end{cases}$$

Example 4: Linearly Constrained Smooth Optimization

$$
\begin{array}{ll}
\min & F(\mathbf{y}) \\
\text{s.t.} & \mathbf{D}\mathbf{y} = \mathbf{b}, \\
& \boldsymbol{\ell} \leq \mathbf{y} \leq \mathbf{u}.
\end{array}
$$

- $\mathbf{D} \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^m$ and $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^d$ satisfy $\boldsymbol{\ell} \leq \mathbf{u}$.
- Fits model (Q) with $G(\mathbf{y}) \equiv \delta_C(\mathbf{y})$,
  $C = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{D}\mathbf{y} = \mathbf{b}, \boldsymbol{\ell} \leq \mathbf{y} \leq \mathbf{u}\}$

Example 4: Linearly Constrained Smooth Optimization

$$\begin{array}{ll} \min & F(\mathbf{y}) \\ \text{s.t.} & \mathbf{D}\mathbf{y} = \mathbf{b}, \\ & \boldsymbol{\ell} \leq \mathbf{y} \leq \mathbf{u}. \end{array}$$

- $\mathbf{D} \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^m$ and $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^d$ satisfy $\boldsymbol{\ell} \leq \mathbf{u}$.
- Fits model (Q) with $G(\mathbf{y}) \equiv \delta_C(\mathbf{y})$,
  $C = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{D}\mathbf{y} = \mathbf{b}, \boldsymbol{\ell} \leq \mathbf{y} \leq \mathbf{u}\}$

Motivation

- **Sparse updates:** Can we find a PDA vector with an appropriate approximation factor, different from $\mathbf{y}$ by only a few components?

Example 4: Linearly Constrained Smooth Optimization

$$\begin{array}{ll} \min & F(\mathbf{y}) \\ \text{s.t.} & \mathbf{D}\mathbf{y} = \mathbf{b}, \\ & \boldsymbol{\ell} \leq \mathbf{y} \leq \mathbf{u}. \end{array}$$

- $\mathbf{D} \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^m$ and $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^d$ satisfy $\boldsymbol{\ell} \leq \mathbf{u}$.
- Fits model (Q) with $G(\mathbf{y}) \equiv \delta_C(\mathbf{y})$, $C = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{D}\mathbf{y} = \mathbf{b}, \boldsymbol{\ell} \leq \mathbf{y} \leq \mathbf{u}\}$

Motivation

- **Sparse updates:** Can we find a PDA vector with an appropriate approximation factor, different from $\mathbf{y}$ by only a few components?
- For $m = 0$ the answer was yes. $\exists \frac{1}{d}$-PDA 1-sparse vector.

Example 4: Linearly Constrained Smooth Optimization

$$\begin{array}{ll} \min & F(\mathbf{y}) \\ \text{s.t.} & \mathbf{D}\mathbf{y} = \mathbf{b}, \\ & \boldsymbol{\ell} \le \mathbf{y} \le \mathbf{u}. \end{array}$$

- $\mathbf{D} \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^m$ and $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^d$ satisfy $\boldsymbol{\ell} \le \mathbf{u}$.
- Fits model (Q) with $G(\mathbf{y}) \equiv \delta_C(\mathbf{y})$,
  $C = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{D}\mathbf{y} = \mathbf{b}, \boldsymbol{\ell} \le \mathbf{y} \le \mathbf{u}\}$

Motivation

- **Sparse updates:** Can we find a PDA vector with an appropriate approximation factor, different from $\mathbf{y}$ by only a few components?
- For $m = 0$ the answer was yes. $\exists \frac{1}{d}$-PDA 1-sparse vector.
- A $\frac{1}{d}$-PDA (m+1)-sparse update vector exists. **Main Idea: LP problems naturally have sparse optimal solutions (bfs's)**
- sparse updates: Platt [99'], Chang et al. [10'], List & Simon [07'], Tseng & Yun [10'] - conformal realizations and review.

The sparseDir Procedure

Input: $\bar{\mathbf{y}} \in C$.
Output: $\mathbf{d}_s(\bar{\mathbf{y}})$ sat. $\|\mathbf{d}_s(\bar{\mathbf{y}})\|_0 \leq m+1$ and $\bar{\mathbf{y}}+\mathbf{d}_s(\bar{\mathbf{y}})$ a $\frac{1}{d}$-PDA vector.

(i) Set

$$\begin{aligned}\mathbf{r} &= \mathbf{p}(\bar{\mathbf{y}}) - \bar{\mathbf{y}},\\ \tilde{\mathbf{D}} &= \mathbf{D}\,\mathrm{diag}(\mathbf{r}),\\ \mathbf{c} &= \mathbf{r} \circ \nabla F(\bar{\mathbf{y}}).\end{aligned}$$

(ii) Compute $\bar{\mathbf{v}}$, a bfs of the linear system s.t. $\langle \mathbf{c}, \bar{\mathbf{v}} \rangle \leq \langle \mathbf{c}, \mathbf{r}^{\dagger} \circ \mathbf{r} \rangle$.

$$\begin{aligned}\tilde{\mathbf{D}}\mathbf{v} &= \mathbf{0},\\ \langle \mathbf{1}, \mathbf{v} \rangle &\leq \|\mathbf{r}\|_0,\\ \mathbf{v} &\geq \mathbf{0}.\end{aligned}$$

(iii) If $\|\mathbf{r}\|_0 = 0$, set $\mathbf{d}_s(\bar{\mathbf{y}}) := \mathbf{0}$. Otherwise set $\mathbf{d}_s(\bar{\mathbf{y}}) := \frac{1}{\|\mathbf{r}\|_0}\mathbf{r} \circ \bar{\mathbf{v}}$.

Example 4: PDA-Based Methods for Linearly Constrained Smooth Minimization

- Variety of methods based Based on the $\frac{1}{d}$-PDA vector $\mathbf{u}_s(\mathbf{y}) \equiv \mathbf{y} + \mathbf{d}_s(\mathbf{y})$.
- Construction of methods depend on the choice of (i) the sets X^k and (ii) the update step (exact/prox-grad).
- First two options fully exploit $\mathbf{u}_s(\mathbf{y}^k)$. The last options only use the following support information:

$$J_k = \{i : \mathbf{u}_s(\mathbf{y}^k)_i = \mathbf{y}_i^k\}.$$

Example 4: PDA-Based Methods for Linearly Constrained Smooth Minimization

- **line segment minimization**
  $(X^k = [\mathbf{y}^k, \mathbf{u}_s(\mathbf{y}^k)],$ exact update$)$
  $$\boxed{\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k),}$$
  where $t_k \in \operatorname{argmin}_{0 \le t \le 1} F(\mathbf{y}^k + t(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k))$.
- **ray minimization**
  $(X^k = \{\mathbf{y}^k + t(\mathbf{u}_s(\mathbf{y}^k) - \mathbf{y}^k) : t \ge 0\},$ exact update$)$ Same update for $\mathbf{y}^{k+1}$, t_k can be as large as possible.
- **block exact minimization**
  $(X^k = \{\mathbf{y} \in C : \mathbf{y}_i = \mathbf{y}_i^k, i \in J_k\},$ exact update$)$
  $$\boxed{\mathbf{y}^{k+1} \in \operatorname{argmin}\{F(\mathbf{y}) : \mathbf{y} \in C, \mathbf{y}_i = \mathbf{y}_i^k, i \in J_k\}.}$$
- **block projected gradient**
  $(X^k = \{\mathbf{y} \in C : \mathbf{y}_i = \mathbf{y}_i^k, i \in J_k\},$ prox-grad update$)$
  $$\boxed{\mathbf{y}^{k+1} = P_{X^k}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla F(\mathbf{y}^k)\right).}$$

Sublinear Rate of Convergence

Theorem.

$$\sum_{i=0}^{k}\lambda_i\left(S(\mathbf{y}^i)-[H(\mathbf{y}^i)-H^*]\right)+H(\mathbf{y}^{k+1})-H^*\le\frac{2\gamma}{k+2\gamma}\left(\frac{2\gamma-2}{k+1}(H(\mathbf{y}^0)-H^*)+C\gamma\right)$$

- γ - approximation factor
- $\lambda_i=\frac{i+2\gamma-1}{\sum_{i=0}^{k}(i+2\gamma-1)}.$, $\boldsymbol{\lambda}\in\Delta_{k+1}$
- $C=$

$$\begin{cases} L\cdot\mathrm{diam}(\text{dom } G)^2, & \text{exact minimization or prox-grad with constant step,}\\ \max\{\eta L,\bar{L}\}\cdot\mathrm{diam}(\text{dom } G)^2, & \text{prox-grad with backtracking.}\end{cases}$$

Sublinear Rate of Convergence

Theorem.

$$\sum_{i=0}^{k}\lambda_i\left(S(\mathbf{y}^i)-[H(\mathbf{y}^i)-H^*]\right)+H(\mathbf{y}^{k+1})-H^*\le\frac{2\gamma}{k+2\gamma}\left(\frac{2\gamma-2}{k+1}(H(\mathbf{y}^0)-H^*)+C\gamma\right)$$

- γ - approximation factor
- $\lambda_i=\frac{i+2\gamma-1}{\sum_{i=0}^{k}(i+2\gamma-1)}$., $\boldsymbol{\lambda}\in\Delta_{k+1}$
- $C=$
$$\begin{cases} L\cdot\mathrm{diam}(\text{dom } G)^2, & \text{exact minimization or} \\ & \text{ prox-grad with constant step,} \\ \max\{\eta L,\bar{L}\}\cdot\mathrm{diam}(\text{dom } G)^2, & \text{prox-grad with backtracking.}\end{cases}$$

Corollary.

$$H(\mathbf{y}^{k+1})-H^*\le\frac{2\gamma}{k+2\gamma}\left(\frac{2\gamma-2}{k+1}(H(\mathbf{y}^0)-H^*)+C\gamma\right).$$

The Dual-Based $\frac{1}{\gamma}$-PDA Method

The Model

$$\text{(P)} \quad \bar{p} \equiv \min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{A}\mathbf{x}) + g(\mathbf{B}\mathbf{x})\}$$

$\mathbf{A} \in \mathbb{R}^{r\times n}$ and $\mathbf{B} \in \mathbb{R}^{q\times n}$.

The Model

$$\text{(P)} \quad \bar{p} \equiv \min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{A}\mathbf{x}) + g(\mathbf{B}\mathbf{x})\}$$

$\mathbf{A} \in \mathbb{R}^{r\times n}$ and $\mathbf{B} \in \mathbb{R}^{q\times n}$.

Assumptions:

(A) $\mathbf{A}$ has full row rank, i.e., $\mathbf{A}\mathbf{A}^T \succ \mathbf{0}$.

(B) $f : \mathbb{R}^r \to \mathbb{R} \cup (-\infty, \infty]$ is proper closed and μ-strongly convex.

(C) $g : \mathbb{R}^q \to \mathbb{R}$ is closed, convex and has a Lipschitz constant L_g.

(D) $\operatorname{dom} g^*$ is closed.

(E) One of the following holds:

  (i) g is polyhedral and $\mathrm{im}(\mathbf{A}^T) \cap \mathbf{B}^T\mathrm{dom}(g^*)$ is nonempty.
  (ii) $\mathrm{im}(\mathbf{A}^T) \cap \mathbf{B}^T\mathrm{ridom}(g^*)$ is nonempty.

The Dual Problem

$$
\text{(D)} \qquad \begin{array}{lll} \bar{q} \equiv & \max & -f^*(\mathbf{w}) - g^*(\mathbf{z}) \\ & \text{s.t.} & \mathbf{A}^T\mathbf{w} + \mathbf{B}^T\mathbf{z} = 0, \\ & & \mathbf{w} \in \mathbb{R}^r, \mathbf{z} \in \mathbb{R}^q. \end{array}
$$

The Dual Problem

$$
\text{(D)} \qquad \begin{array}{lll} \bar{q} \equiv & \max & -f^*(\mathbf{w}) - g^*(\mathbf{z}) \\ & \text{s.t.} & \mathbf{A}^T\mathbf{w} + \mathbf{B}^T\mathbf{z} = 0, \\ & & \mathbf{w} \in \mathbb{R}^r, \mathbf{z} \in \mathbb{R}^q. \end{array}
$$

Properties:

- $f^* : \mathbb{R}^r \to \mathbb{R}$ convex and $\frac{1}{\mu}$-smooth.
- $g^* : \mathbb{R}^q \to (\infty, \infty]$ proper closed and convex, $\mathrm{dom}(g^*) \subseteq B[\mathbf{0}, L_g]$.
- If (E.i) is satisfied, then g^* is also polyhedral and $\mathrm{dom}\, g^*$ is a polytope.
- The feasible set

$$
X \equiv \left\{(\mathbf{w}, \mathbf{z}) : \mathbf{z} \in \mathrm{dom}(g^*), \mathbf{A}^T\mathbf{w} + \mathbf{B}^T\mathbf{z} = 0\right\}
$$

  is compact.
- The optimal values, $\bar{p}$ and $\bar{q}$, of problems (P) and (D) are finite, attained and equal.

Reduction of the Dual Problem

- The dual problem (D) can be reduced to

$$\text{(D')} \quad \min_{\mathbf{z}\in\mathbb{R}^q}\{H_1(\mathbf{z}) \equiv F_1(\mathbf{z}) + G_1(\mathbf{z})\}.$$

- Problem (D') fits the general model (Q) with

$$\begin{aligned} F(\mathbf{z}) &= F_1(\mathbf{z}) \equiv f^*(-(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}), \\ G(\mathbf{z}) &= G_1(\mathbf{z}) \equiv g^*(\mathbf{z}) + \delta_{\{\mathbf{p}:(\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{p}=\mathbf{0}\}}(\mathbf{z}). \end{aligned}$$

  where $\mathbf{P} \equiv \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$.

Dual-Based $\frac{1}{\gamma}$-PDA Method

Initialization. $\mathbf{z}^0$ satisfying $(\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{z}^0 = \mathbf{0}, \mathbf{z}^0 \in \operatorname{dom} g^*$.
General Step. For $k = 0, 1, 2 \ldots,$

(i)
- Choose $\mathbf{u}(\mathbf{z}^k)$- a $\frac{1}{\gamma}$-PDA vector of H_1 at $\mathbf{z}^k$.
- Choose a compact set Z^k for which $[\mathbf{z}^k, \mathbf{u}(\mathbf{z}^k)] \subseteq Z^k$.

Dual-Based $\frac{1}{\gamma}$-PDA Method

Initialization. $\mathbf{z}^0$ satisfying $(\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{z}^0 = \mathbf{0}, \mathbf{z}^0 \in \operatorname{dom} g^*$.
General Step. For $k = 0, 1, 2 \ldots,$

(i)
- Choose $\mathbf{u}(\mathbf{z}^k)$- a $\frac{1}{\gamma}$-PDA vector of H_1 at $\mathbf{z}^k$.
- Choose a compact set Z^k for which $[\mathbf{z}^k, \mathbf{u}(\mathbf{z}^k)] \subseteq Z^k$.

(ii) Perform one of the following:
Prox-grad update: $\mathbf{z}^{k+1} = \operatorname{prox}_{\frac{1}{L_k}G_1+\delta_{Z^k}}\left(\mathbf{z}^k - 1/L_k \nabla F_1(\mathbf{z}^k)\right)$
Exact update: $\mathbf{z}^{k+1} = \underset{\mathbf{z}\in Z^k}{\operatorname{argmin}}\, F_1(\mathbf{z}) + G_1(\mathbf{z})$

Dual-Based $\frac{1}{\gamma}$-PDA Method

Initialization. $\mathbf{z}^0$ satisfying $(\mathbf{I}-\mathbf{P})\mathbf{B}^T\mathbf{z}^0 = \mathbf{0}, \mathbf{z}^0 \in \operatorname{dom} g^*$.
General Step. For $k = 0, 1, 2 \ldots,$

(i)
- Choose $\mathbf{u}(\mathbf{z}^k)$- a $\frac{1}{\gamma}$-PDA vector of H_1 at $\mathbf{z}^k$.
- Choose a compact set Z^k for which $[\mathbf{z}^k, \mathbf{u}(\mathbf{z}^k)] \subseteq Z^k$.

(ii) Perform one of the following:
Prox-grad update: $\mathbf{z}^{k+1} = \operatorname{prox}_{\frac{1}{L_k}G_1+\delta_{Z^k}}\left(\mathbf{z}^k - 1/L_k \nabla F_1(\mathbf{z}^k)\right)$
Exact update: $\mathbf{z}^{k+1} = \underset{\mathbf{z}\in Z^k}{\operatorname{argmin}}\, F_1(\mathbf{z}) + G_1(\mathbf{z})$

(iii) Set $\mathbf{w}^k = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}^k$ and compute $\mathbf{s}^k$ by either:

Averaging: $\mathbf{s}^k = \dfrac{1}{\sum_{i=0}^k (i+2\gamma-1)} \displaystyle\sum_{i=0}^k (i+2\gamma-1)\nabla f^*(\mathbf{w}^i)$

Best iterate: $\mathbf{s}^k = \nabla f^*(\mathbf{w}^{k_0}),\ k_0 \in \underset{i=0,1,\ldots k}{\operatorname{argmin}} \{S_1(\mathbf{z}^i) - H_1(\mathbf{z}^i)\}$

(iv) Compute $\mathbf{x}^k \in \underset{\mathbf{x}}{\operatorname{argmin}} \left\{g(\mathbf{B}\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{s}^k\right\}$

Main Convergence Result

Key technical result:

Lemma.

$$S_1(\mathbf{z}) = \min_{\mathbf{A}\mathbf{x}=\nabla f^*(\mathbf{w})} g(\mathbf{B}\mathbf{x}) + f(\nabla f^*(\mathbf{w})) + g^*(\mathbf{z}) + f^*(\mathbf{w})$$

with $\mathbf{w} = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}$.

Main Convergence Result

Key technical result:

Lemma.

$$S_1(\mathbf{z}) = \min_{\mathbf{A}\mathbf{x}=\nabla f^*(\mathbf{w})} g(\mathbf{B}\mathbf{x}) + f(\nabla f^*(\mathbf{w})) + g^*(\mathbf{z}) + f^*(\mathbf{w})$$

with $\mathbf{w} = -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\mathbf{z}$.

$+$ convergence result for the primal PDA method $\Rightarrow$

Theorem (primal-dual convergence) $\mathbf{z}^k$ is dual feasible, $\mathbf{x}^k$ is primal feasible and

$$f(\mathbf{A}\mathbf{x}^k)+g(\mathbf{B}\mathbf{x}^k)+H_1(\mathbf{z}^{k+1}) \le \frac{2\gamma}{k+2\gamma}\left(\frac{2\gamma-2}{k+1}(H_1(\mathbf{z}^0)-\bar{p})+4\tilde{C}\gamma\right),$$

$$\tilde{C} = \begin{cases} \frac{\|(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\|^2}{\mu}L_g^2, & \text{exact min., prox-grad with constant step,} \\ \max\left\{\eta\frac{\|(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{B}^T\|^2}{\mu}, \bar{L}\right\}L_g^2, & \text{prox-grad, backtracking.} \end{cases}$$

Example 1: Binary Classification with Offset(SVM)

- **Given.** q datapoints $(\mathbf{s}_i, t_i)$, where $\mathbf{s}_i \in \mathbb{R}^n$ are the feature vectors $t_i \in \{-1, 1\}$ are the binary outputs.
- **Objective.** find a pair $(\mathbf{x}, b) \in \mathbb{R}^n \times \mathbb{R}$ such that the hyperplane $\{\mathbf{y} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{y} \rangle = b\}$ will be a "good" separator between the two types of datapoints.

Example 1: Binary Classification with Offset(SVM)

- **Given.** q datapoints $(\mathbf{s}_i, t_i)$, where $\mathbf{s}_i \in \mathbb{R}^n$ are the feature vectors $t_i \in \{-1, 1\}$ are the binary outputs.
- **Objective.** find a pair $(\mathbf{x}, b) \in \mathbb{R}^n \times \mathbb{R}$ such that the hyperplane $\{\mathbf{y} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{y} \rangle = b\}$ will be a "good" separator between the two types of datapoints.
- **Model.** minimize a penalized empirical risk.

$$\min_{\mathbf{x}, b} \frac{1}{2}\|\mathbf{x}\|^2 + \frac{C}{q}\sum_{i=1}^{q} \ell(t_i(\langle \mathbf{x}, \mathbf{s}_i \rangle - b))$$

  - $C > 0$ – regularization parameter.
  - $\ell : \mathbb{R} \to \mathbb{R}$ – convex Lipschitz and nonincreasing loss function.

popular choice for l: $\ell(z) = \max\{1 - z, 0\}$ – hinge loss.

Binary Classification Contd.

$$\boxed{(1) \min_{\mathbf{x},b} \frac{1}{2}\|\mathbf{x}\|^2 + \frac{C}{q}\sum_{i=1}^q \ell(t_i(\langle \mathbf{x}, \mathbf{s}_i\rangle - b))}$$

$$\boxed{(\mathrm{P}) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{A}\mathbf{x}) + g(\mathbf{B}\mathbf{x})\}}$$

Problem (1) fits model (P) with

- $f(\mathbf{w}) \equiv \frac{1}{2}\|\mathbf{w}\|^2$
- $g(\mathbf{u}) \equiv \frac{C}{q}\sum_{i=1}^q \ell(u_i)$
- $\mathbf{A} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1}\end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{S}^T & -\mathbf{t}\end{pmatrix}$

Binary Classification Contd.

$$\boxed{(1)\ \min_{\mathbf{x},b} \tfrac{1}{2}\|\mathbf{x}\|^2 + \tfrac{C}{q}\textstyle\sum_{i=1}^q \ell(t_i(\langle \mathbf{x}, \mathbf{s}_i\rangle - b))}$$

$$\boxed{(\mathrm{P}) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{A}\mathbf{x}) + g(\mathbf{B}\mathbf{x})\}}$$

Problem (1) fits model (P) with

- $f(\mathbf{w}) \equiv \frac{1}{2}\|\mathbf{w}\|^2$
- $g(\mathbf{u}) \equiv \frac{C}{q}\sum_{i=1}^q \ell(u_i)$
- $\mathbf{A} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1}\end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{S}^T & -\mathbf{t}\end{pmatrix}$

Problem (D) then becomes

$$(2) \quad \begin{array}{ll} \min & \frac{1}{2}\mathbf{z}^T\mathbf{S}^T\mathbf{S}\mathbf{z} + \frac{C}{q}\sum_{i=1}^q \ell^*\left(\frac{qz_i}{C}\right) \\ \text{s.t.} & \mathbf{t}^T\mathbf{z} = 0. \end{array}$$

Binary Classification Contd.

$$(1)\ \min_{\mathbf{x},b} \tfrac{1}{2}\|\mathbf{x}\|^2 + \tfrac{C}{q}\sum_{i=1}^q \ell(t_i(\langle \mathbf{x}, \mathbf{s}_i\rangle - b))$$

$$(\mathrm{P}) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{A}\mathbf{x}) + g(\mathbf{B}\mathbf{x})\}$$

Problem (1) fits model (P) with

- $f(\mathbf{w}) \equiv \frac{1}{2}\|\mathbf{w}\|^2$
- $g(\mathbf{u}) \equiv \frac{C}{q}\sum_{i=1}^q \ell(u_i)$
- $\mathbf{A} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1}\end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{S}^T & -\mathbf{t}\end{pmatrix}$

Problem (D) then becomes

$$(2) \begin{array}{ll} \min & \frac{1}{2}\mathbf{z}^T\mathbf{S}^T\mathbf{S}\mathbf{z} + \frac{C}{q}\sum_{i=1}^q \ell^*\left(\frac{qz_i}{C}\right) \\ \text{s.t.} & \mathbf{t}^T\mathbf{z} = 0. \end{array}$$

If ℓ is the hinge loss function,

$$\begin{array}{ll} \min & \frac{1}{2}\mathbf{z}^T\mathbf{S}^T\mathbf{S}\mathbf{z} + \mathbf{1}^T\mathbf{z}, \\ \text{s.t.} & -\frac{C}{q} \le z_i \le 0, i = 1, 2, \ldots, q \\ & \mathbf{t}^T\mathbf{z} = 0, \end{array}$$

Binary Classification Contd.

- can use the $\frac{1}{q}$-PDA method on the dual problem described before (also amounts to a 2-CD method)
- working set choice is done in $O(q)$ flops (solution fractional knapsack problem).
- the step that is done after choosing the 2 coordinates can be done by either exact minimization, conditional gradient, gradient projection,
- [Hush et al. 2006'] have a $O(1/\sqrt{k})$ rate of convergence result for the primal sequence.

Numerical Simulations

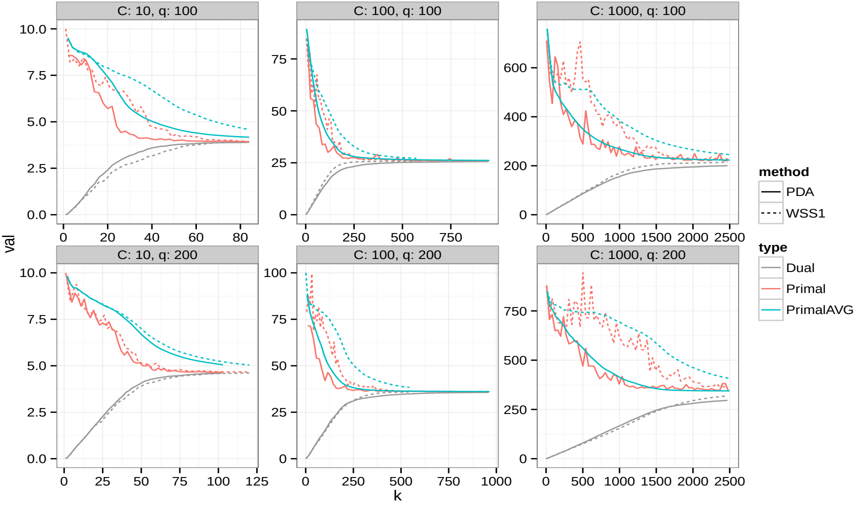
Setting

- The ambient dimension is $p = 20$.
- two classes sampled from unit Gaussian random variables with a shift in mean of magnitude 2.
- number of datapoints $q \in \{100, 200\}$.
- regularization parameter $C \in \{10, 100, 1000\}$.

Methods

- PDA – the $\frac{1}{q}$-dual based method with exact minimization step.
- WSS1 – from LIBSVM. Identical to PDA, but with a different index selection rule.

Numerical Results



has better performance than WSS1, averaging doesn't seem to help in this problem.

THANK YOU FOR YOUR ATTENTION!

- A. Beck, E. Pauwels and S. Sabach, "Primal and dual predicted decrease approximation methods", Mathematical Programming (2017).