Generalized Stochastic Frank-Wolfe Algorithm with Stochastic "Substitute" Gradient for Structured Convex Optimization

Haihao (Sean) Lu, Robert M. Freund

MIT

INFORMS Phoenix, November 2018

Paper on arXiv (and in review):

"Generalized Stochastic Frank-Wolfe Algorithm with Stochastic "Substitute" Gradient for Structured Convex Optimization"

Overview/Results

- Introduction
  - Problem of Interest
  - Examples in Statistical and Machine Learning
  - Literature Review
  - Primal-Dual Structure
- Stochastic Generalized Frank-Wolfe and Randomized Dual Coordinate Mirror Descent
  - Substitute Gradient
  - Stochastic Generalized Frank-Wolfe (SGFW)
  - Randomized Dual Coordinate Mirror Descent (RDCMD)
  - Equivalence of SGFW and RDCMD
- Computational Guarantees of SGFW and RDCMD
  - $O(1/\varepsilon)$ Sublinear Convergence Rate
    - First-Order Methods Naturally Minimize a Primal-Dual Gap
    - Randomized Coordinate Descent for Nonsmooth Functions
  - Linear Convergence when the regularizer is Strongly Convex
  - Extensions/Discussions
- Contributions/Summary

Problem of Interest

The problem of interest is

$$\textbf{P}: \qquad \min_{\beta} P(\beta) := \frac{1}{n}\sum_{j=1}^{n} l_j(x_j^T\beta) + R(\beta) \ ,$$

- $l_j(\cdot)$ is a univariate loss function
- $R(\cdot)$ is a regularizer and/or an indicator function of a feasible region Q and/or a penalty term, coupling constraints, etc.
- In standard Frank-Wolfe setting, $R(\cdot)$ is an indicator function

Assuptions

Assumptions

1. For $j = 1, \ldots, n$, the univariate function $l_j(\cdot)$ is strictly convex and γ-smooth, namely for all a and b,

$$|\dot{l}_j(a) - \dot{l}_j(b)| \leq \gamma |a - b|$$

2. $\mathrm{dom} R(\cdot)$ is bounded, and the subproblem

$$\min_{\beta} c^T \beta + R(\beta)$$

attains its optimum and can be easily solved for any c

3. $0 \in \mathrm{dom} R(\cdot)$

Examples in Statistical and Machine Learning

- LASSO

$$\min_\beta \quad \frac{1}{2n}\sum_{j=1}^n (y_j - x_j^T\beta)^2$$

$$s.t. \quad \|\beta\|_1 \le \delta \ ,$$

where $l_j(\cdot) = \frac{1}{2}(y_j - \cdot)^2$ and $R(\beta) := \mathbf{I}_{\{\|\beta\|_1 \le \delta\}}(\beta)$
(Here $\mathbf{I}_Q(\cdot)$ is the indicator function on the set Q.)

- Sparse Logistic Regression

$$\min_\beta \frac{1}{n}\sum_{j=1}^n \ln(1+\exp(-y_j x_j^T\beta)) + \lambda\|\beta\|_1 \ ,$$

where $l_j(\cdot) = \ln(1+\exp(-y_j\cdot))$, $R(\beta) = \lambda\|\beta\|_1 + \mathbf{I}_{\{\|\beta\|_1 \le \ln(2)/\lambda\}}(\beta)$

- Matrix Completion

$$\min_{\beta\in\mathbb{R}^{n\times p}} \quad \frac{1}{2|\Omega|}\sum_{(i,j)\in\Omega}(M_{i,j} - \beta_{i,j})^2$$

$$s.t. \quad \|\beta\|_* \le \delta \ ,$$

where $l_{(i,j)}(\cdot) = \frac{1}{2}(\cdot - M_{i,j})^2$ and $R(\beta) = \mathbf{I}_{\{\|\beta\|_* \le \delta\}}(\beta)$

- More examples can be found in [Jaggi 2013].

Frank-Wolfe and Generalized Frank-Wolfe

In the traditional Frank-Wolfe setting $R(\cdot)$ is an indicator function of a bounded set Q, and the Frank-Wolfe update is:

Traditional Frank-Wolfe Method

$$\tilde{\beta}^i \in \arg\min_{\beta \in Q} \left\{ \nabla f(\beta^i)^T \beta \right\} \quad \text{and} \quad \beta^{i+1} = (1-\alpha_i)\beta^i + \alpha_i \tilde{\beta}^i$$

In the generalized Frank-Wolfe setting where $R(\cdot)$ can be any convex function, the Generalized Frank-Wolfe update is:

Generalized Frank-Wolfe Method

$$\tilde{\beta}^i \in \arg\min \left\{ \nabla f(\beta^i)^T \beta + R(\beta) \right\} \quad \text{and} \quad \beta^{i+1} = (1-\alpha_i)\beta^i + \alpha_i \tilde{\beta}^i$$

Stochastic Frank-Wolfe Method

In the stochastic setting, we can only compute an unbiased estimator $\tilde{g}^i$ of the gradient $\nabla f(\beta^i)$, and the update is

Stochastic Frank-Wolfe Method

$$\tilde{\beta}^i \in \arg\min_{\beta \in Q} \left\{ (\tilde{g}^i)^T \beta \right\} \quad \text{and} \quad \beta^{i+1} = (1-\alpha_i)\beta^i + \alpha_i \tilde{\beta}^i$$

Stochastic Frank-Wolfe Method

Algorithm and Reference	Number of Exact Gradient Calls	Number of Stochastic Gradient Calls	Number of Linear Optimization Oracle Calls
FW*	$O(\frac{1}{\varepsilon})$	0	$O(\frac{1}{\varepsilon})$
SFW**	0	$O(\frac{1}{\varepsilon^3})$	$O(\frac{1}{\varepsilon})$
Online-FW***	0	$O(\frac{1}{\varepsilon^4})$	$O(\frac{1}{\varepsilon^4})$
SCGS****	0	$O(\frac{1}{\varepsilon^2})$	$O(\frac{1}{\varepsilon})$
SVRFW**	$O(\ln \frac{1}{\varepsilon})$	$O(\frac{1}{\varepsilon^2})$	$O(\frac{1}{\varepsilon})$
STORC**	$O(\ln \frac{1}{\varepsilon})$	$O(\frac{1}{\varepsilon^{1.5}})$	$O(\frac{1}{\varepsilon})$
This work	1	$O(\frac{1}{\varepsilon})$	$O(\frac{1}{\varepsilon})$

*[Frank, Wolfe 1956], **[Hazan, Luo 2016], ***[Hazan, Kale 2012], ****[Lan, Zhou 2016]

Conjugate Function

Recall the definition of the conjugate of a function $f(\cdot)$:

$$f^*(y) := \sup_{x \in \mathrm{dom} f(\cdot)} \{y^T x - f(x)\} \ .$$

Proposition: Conjugate Functions

If $f(\cdot)$ is a closed convex function, then $f^{**}(\cdot) = f(\cdot)$. Furthermore:

1. $f(\cdot)$ is γ-smooth with domain $\mathbb{R}^p$ with respect to the norm $\|\cdot\|$ if and only if $f^*(\cdot)$ is $1/\gamma$-strongly convex with respect to the (dual) norm $\|\cdot\|^*$.
2. If $f(\cdot)$ is differentiable and strictly convex, then the following three conditions are equivalent:
   - $y = \nabla f(x)$
   - $x = \nabla f^*(y)$, and
   - $x^T y = f(x) + f^*(y)$.

Primal-Dual Structure

The original problem is

$$\textbf{P}: \qquad \min_{\beta} P(\beta) := \tfrac{1}{n}\sum_{j=1}^{n} l_j(x_j^T\beta) + R(\beta) \ .$$

Denote $X := [x_1^T; x_2^T; \ldots; x_n^T]$. Then the corresponding dual problem is

$$\textbf{D}: \qquad \max_{w} D(w) := -R^*\left(-\tfrac{1}{n}X^T w\right) - \tfrac{1}{n}\sum_{j=1}^{n} l_j^*(w_j) \ .$$

Define the convex/concave saddle-function $\phi(\cdot,\cdot)$:

$$\phi(\beta, w) := \tfrac{1}{n} w^T X\beta - \tfrac{1}{n}\sum_{i=1}^{n} l_i^*(w_i) + R(\beta) \ .$$

We can write P and D in saddlepoint minimax format as:

$$\textbf{P}: \ \min_{\beta}\max_{w} \phi(\beta, w) \qquad \text{and} \qquad \textbf{D}: \ \max_{w}\min_{\beta} \phi(\beta, w) \ .$$

SGFW and RDCM

Stochastic Generalized Frank-Wolfe

and

Randomized Dual Coordinate Mirror Descent

"Substitute" Gradient

The problem of interest is

$$\mathbf{P}: \qquad \min_{\beta} P(\beta) := \frac{1}{n} \sum_{j=1}^{n} l_j(x_j^T \beta) + R(\beta) \ .$$

The gradient of the first term is

$$\tfrac{1}{n} \sum_{j=1}^{n} \dot{l}_j(x_j^T \beta) x_j = \tfrac{1}{n} \sum_{j=1}^{n} \dot{l}_j(s_j) x_j \quad \text{where} \quad s_j = x_j^T \beta$$

It is too expensive to update $x_j^T \beta$ for all $j = 1, \ldots, n$ in each iteration when n is large. "Substitute" gradient d is computed by

$$d = \tfrac{1}{n} \sum_{j=1}^{n} \dot{l}_j(s_j) x_j, \ j = 1, \ldots, n \ .$$

- We will only update one s_j in each iteration
- As a result d will not in general be an unbiased estimator of the gradient

Stochastic Generalized Frank-Wolfe Method with Substitute Gradient

Stochastic Generalized Frank-Wolfe with Substitute Gradient(SGFW)

Initialize with $\bar{\beta}^{-1} = 0$, $s^0 = 0$, and substitute gradient $d^0 = \frac{1}{n} X^T \nabla L(s^0)$, with step-size sequences $\{\alpha_i\} \in (0,1]$, $\{\eta_i\} \in (0,1]$.

For iterations $i = 0, 1, \ldots,$ do:

Solve l.o.o. subproblem: Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(d^i\right)^T \beta + R(\beta) \right\}$

Choose random index: Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Update s value: $s_{j_i}^{i+1} \leftarrow (1-\eta_i) s_{j_i}^i + \eta_i (x_{j_i}^T \tilde{\beta}^i)$, and $s_j^{i+1} \leftarrow s_j^i$ for $j \neq j_i$

Update substitute gradient:

$$d^{i+1} = \frac{1}{n} X^T \nabla L(s^{i+1}) = d^i + \frac{1}{n} \left(\dot{l}_{j_i}(s_{j_i}^{i+1}) - \dot{l}_{j_i}(s_{j_i}^i) \right) x_{j_i}$$

Update primal variable: $\bar{\beta}^i \leftarrow (1-\alpha_i) \bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

(Optional Accounting:) $w^{i+1} \leftarrow \nabla L(s^{i+1})$

Stochastic Generalized Frank-Wolfe Method with Substitute Gradient

Stochastic Generalized Frank-Wolfe with Substitute Gradient(SGFW)

Initialize with $\bar{\beta}^{-1} = 0$, $s^0 = 0$, and substitute gradient $d^0 = \frac{1}{n} X^T \nabla L(s^0)$, with step-size sequences $\{\alpha_i\} \in (0, 1]$, $\{\eta_i\} \in (0, 1]$.

For iterations $i = 0, 1, \ldots,$ do:

Solve l.o.o. subproblem: Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(d^i\right)^T \beta + R(\beta) \right\}$

Choose random index: Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Update s value: $s_{j_i}^{i+1} \leftarrow (1 - \eta_i) s_{j_i}^i + \eta_i (x_{j_i}^T \tilde{\beta}^i)$, and $s_j^{i+1} \leftarrow s_j^i$ for $j \neq j_i$

Update substitute gradient:
$d^{i+1} = \frac{1}{n} X^T \nabla L(s^{i+1}) = d^i + \frac{1}{n} \left(\dot{l}_{j_i}(s_{j_i}^{i+1}) - \dot{l}_{j_i}(s_{j_i}^i) \right) x_{j_i}$

Update primal variable: $\bar{\beta}^i \leftarrow (1 - \alpha_i) \bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

(Optional Accounting:) $w^{i+1} \leftarrow \nabla L(s^{i+1})$

Stochastic Generalized Frank-Wolfe Method with Substitute Gradient

Stochastic Generalized Frank-Wolfe with Substitute Gradient(SGFW)

Initialize with $\bar{\beta}^{-1} = 0$, $s^0 = 0$, and substitute gradient $d^0 = \frac{1}{n} X^T \nabla L(s^0)$, with step-size sequences $\{\alpha_i\} \in (0, 1]$, $\{\eta_i\} \in (0, 1]$.

For iterations $i = 0, 1, \ldots,$ do:

Solve l.o.o. subproblem: Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(d^i\right)^T \beta + R(\beta) \right\}$

Choose random index: Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Update s value: $s_{j_i}^{i+1} \leftarrow (1 - \eta_i) s_{j_i}^i + \eta_i (x_{j_i}^T \tilde{\beta}^i)$, and $s_j^{i+1} \leftarrow s_j^i$ for $j \neq j_i$

Update substitute gradient:

$$d^{i+1} = \frac{1}{n} X^T \nabla L(s^{i+1}) = d^i + \frac{1}{n} \left(\dot{l}_{j_i}(s_{j_i}^{i+1}) - \dot{l}_{j_i}(s_{j_i}^i) \right) x_{j_i}$$

Update primal variable: $\bar{\beta}^i \leftarrow (1 - \alpha_i) \bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

(Optional Accounting:) $w^{i+1} \leftarrow \nabla L(s^{i+1})$

Stochastic Generalized Frank-Wolfe Method with Substitute Gradient

Stochastic Generalized Frank-Wolfe with Substitute Gradient(SGFW)

Initialize with $\bar{\beta}^{-1} = 0$, $s^0 = 0$, and substitute gradient $d^0 = \frac{1}{n} X^T \nabla L(s^0)$, with step-size sequences $\{\alpha_i\} \in (0, 1]$, $\{\eta_i\} \in (0, 1]$.

For iterations $i = 0, 1, \ldots,$ do:

Solve l.o.o. subproblem: Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(d^i\right)^T \beta + R(\beta) \right\}$

Choose random index: Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Update s value: $s_{j_i}^{i+1} \leftarrow (1 - \eta_i) s_{j_i}^i + \eta_i (x_{j_i}^T \tilde{\beta}^i)$, and $s_j^{i+1} \leftarrow s_j^i$ for $j \neq j_i$

Update substitute gradient:

$$d^{i+1} = \frac{1}{n} X^T \nabla L(s^{i+1}) = d^i + \frac{1}{n} \left(\dot{l}_{j_i}(s_{j_i}^{i+1}) - \dot{l}_{j_i}(s_{j_i}^i) \right) x_{j_i}$$

Update primal variable: $\bar{\beta}^i \leftarrow (1 - \alpha_i) \bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

(Optional Accounting:) $w^{i+1} \leftarrow \nabla L(s^{i+1})$

Stochastic Generalized Frank-Wolfe Method with Substitute Gradient

Stochastic Generalized Frank-Wolfe with Substitute Gradient(SGFW)

Initialize with $\bar{\beta}^{-1} = 0$, $s^0 = 0$, and substitute gradient $d^0 = \frac{1}{n} X^T \nabla L(s^0)$, with step-size sequences $\{\alpha_i\} \in (0, 1]$, $\{\eta_i\} \in (0, 1]$.

For iterations $i = 0, 1, \ldots,$ do:

Solve l.o.o. subproblem: Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(d^i\right)^T \beta + R(\beta) \right\}$

Choose random index: Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Update s value: $s_{j_i}^{i+1} \leftarrow (1 - \eta_i) s_{j_i}^i + \eta_i (x_{j_i}^T \tilde{\beta}^i)$, and $s_j^{i+1} \leftarrow s_j^i$ for $j \neq j_i$

Update substitute gradient:

$d^{i+1} = \frac{1}{n} X^T \nabla L(s^{i+1}) = d^i + \frac{1}{n} \left(\dot{l}_{j_i}(s_{j_i}^{i+1}) - \dot{l}_{j_i}(s_{j_i}^i) \right) x_{j_i}$

Update primal variable: $\bar{\beta}^i \leftarrow (1 - \alpha_i) \bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

(Optional Accounting:) $w^{i+1} \leftarrow \nabla L(s^{i+1})$

Stochastic Generalized Frank-Wolfe Method with Substitute Gradient

Stochastic Generalized Frank-Wolfe with Substitute Gradient (SGFW)

Initialize with $\bar{\beta}^{-1} = 0$, $s^0 = 0$, and substitute gradient $d^0 = \frac{1}{n} X^T \nabla L(s^0)$, with step-size sequences $\{\alpha_i\} \in (0, 1]$, $\{\eta_i\} \in (0, 1]$.

For iterations $i = 0, 1, \ldots,$ do:

Solve l.o.o. subproblem: Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(d^i\right)^T \beta + R(\beta) \right\}$

Choose random index: Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Update s value: $s_{j_i}^{i+1} \leftarrow (1 - \eta_i) s_{j_i}^i + \eta_i (x_{j_i}^T \tilde{\beta}^i)$, and $s_j^{i+1} \leftarrow s_j^i$ for $j \neq j_i$

Update substitute gradient:

$$d^{i+1} = \frac{1}{n} X^T \nabla L(s^{i+1}) = d^i + \frac{1}{n} \left(\dot{l}_{j_i}(s_{j_i}^{i+1}) - \dot{l}_{j_i}(s_{j_i}^i) \right) x_{j_i}$$

Update primal variable: $\bar{\beta}^i \leftarrow (1 - \alpha_i) \bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$

(Optional Accounting:) $w^{i+1} \leftarrow \nabla L(s^{i+1})$

Stochastic Generalized Frank-Wolfe Method with Substitute Gradient

Stochastic Generalized Frank-Wolfe with Substitute Gradient (SGFW)

Initialize with $\bar{\beta}^{-1} = 0$, $s^0 = 0$, and substitute gradient $d^0 = \frac{1}{n} X^T \nabla L(s^0)$, with step-size sequences $\{\alpha_i\} \in (0,1]$, $\{\eta_i\} \in (0,1]$.

For iterations $i = 0, 1, \ldots,$ do:

Solve l.o.o. subproblem: Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(d^i\right)^T \beta + R(\beta) \right\}$

Choose random index: Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Update s value: $s_{j_i}^{i+1} \leftarrow (1 - \eta_i) s_{j_i}^i + \eta_i (x_{j_i}^T \tilde{\beta}^i)$, and $s_j^{i+1} \leftarrow s_j^i$ for $j \neq j_i$

Update substitute gradient:

$$d^{i+1} = \frac{1}{n} X^T \nabla L(s^{i+1}) = d^i + \frac{1}{n} \left(\dot{l}_{j_i}(s_{j_i}^{i+1}) - \dot{l}_{j_i}(s_{j_i}^i) \right) x_{j_i}$$

Update primal variable: $\bar{\beta}^i \leftarrow (1 - \alpha_i) \bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$

(Optional Accounting:) $w^{i+1} \leftarrow \nabla L(s^{i+1})$

Stochastic Generalized Frank-Wolfe Method with Substitute Gradient

Remarks

- SGFW takes place completely in the primal space
- We used two step-size sequences:
  - $\{\eta_i\}$ is used to update the s_{j_i} values
  - $\{\alpha_i\}$ is used to update the $\bar{\beta}^i$ values

Randomized Dual Coordinate Mirror Descent

The Dual Problem

$$\max_{w} D(w) := -R^* \left(-\tfrac{1}{n} X^T w\right) - \tfrac{1}{n} \sum_{j=1}^{n} l_j^*(w_j) \ .$$

- $D(w)$ may not be differentiable, but it is strongly convex.
- Let us define $L^*(w) := \sum_{j=1}^{n} l_j^*(w_j)$ and

$$\tilde{\beta}^i \in \arg\min_{\beta} \left\{ \left(\tfrac{1}{n}(w^i)^T X\beta + R(\beta)\right) \right\} \ ,$$

then it turns out

$$g^i := \frac{1}{n} \left(X\tilde{\beta}^i - \nabla L^*(w^i) \right) \in \partial D(w^i) \ .$$

- Therefore

$$\tilde{g}^i \leftarrow \frac{1}{n} \left(x_{j_i}^T \tilde{\beta}^i - \dot{l}_{j_i}^*(w_{j_i}^i) \right) e_{j_i}$$

is a coordinate of a subgradient of $D(w)$ at w^i.

Randomized Dual Coordinate Mirror Descent

Randomized Dual Coordinate Mirror Descent (RDCMD)

Define the prox function $h(w) := \frac{1}{n}\sum_{j=1}^{n} l_j^*(w_j)$. Initialize with $w^0 = \arg\min_w \frac{1}{n}\sum_{j=1}^{n} l_j^*(w_j)$ and step-size sequences $\{\alpha_i\} \in (0,1]$ and $\{\eta_i\} \in (0,1]$. (Optional: set $\bar{\beta}^{-1} = 0$.)

For iterations $i = 0, 1, \ldots$

Compute Randomized Coordinate of Subgradient of $D(\cdot)$ at w^i
Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(\frac{1}{n}(w^i)^T X\beta + R(\beta) \right) \right\}$

Choose random index. Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Compute subgradient coordinate vector: $\tilde{g}^i \leftarrow \frac{1}{n}\left(x_{j_i}^T \tilde{\beta}^i - \dot{l}_{j_i}^*(w_{j_i}^i) \right) e_{j_i}$

Update dual variable: Compute
$w^{i+1} = \arg\min_w \left\{ \left\langle -\eta_i \tilde{g}^i, w - w^i \right\rangle + D_h(w, w^i) \right\}$

(Optional Accounting:) $\bar{\beta}^i \leftarrow (1-\alpha_i)\bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

Randomized Dual Coordinate Mirror Descent

Randomized Dual Coordinate Mirror Descent (RDCMD)

Define the prox function $h(w) := \frac{1}{n}\sum_{i=1}^{n} l_i^*(w_i)$. Initialize with $w^0 = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} l_i^*(w_i)$ and step-size sequences $\{\alpha_i\} \in (0,1]$ and $\{\eta_i\} \in (0,1]$. (Optional: set $\bar{\beta}^{-1} = 0$.)

For iterations $i = 0, 1, \ldots$

Compute Randomized Coordinate of Subgradient of $D(\cdot)$ **at** w^i

Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(\frac{1}{n}(w^i)^T X\beta + R(\beta) \right) \right\}$

Choose random index. Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Compute subgradient coordinate vector: $\tilde{g}^i \leftarrow \frac{1}{n}\left(x_{j_i}^T \tilde{\beta}^i - \dot{l}_{j_i}^*(w_{j_i}^i) \right) e_{j_i}$

Update dual variable: Compute
$w^{i+1} = \arg\min_w \left\{ \left\langle -\eta_i \tilde{g}^i, w - w^i \right\rangle + D_h(w, w^i) \right\}$

(Optional Accounting:) $\bar{\beta}^i \leftarrow (1-\alpha_i)\bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

Randomized Dual Coordinate Mirror Descent

Randomized Dual Coordinate Mirror Descent (RDCMD)

Define the prox function $h(w) := \frac{1}{n}\sum_{i=1}^{n} l_i^*(w_i)$. Initialize with $w^0 = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} l_i^*(w_i)$ and step-size sequences $\{\alpha_i\} \in (0,1]$ and $\{\eta_i\} \in (0,1]$. (Optional: set $\bar{\beta}^{-1} = 0$.)

For iterations $i = 0, 1, \ldots$

Compute Randomized Coordinate of Subgradient of $D(\cdot)$ **at** w^i
Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(\frac{1}{n}(w^i)^T X\beta + R(\beta) \right) \right\}$

Choose random index. Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Compute subgradient coordinate vector: $\tilde{g}^i \leftarrow \frac{1}{n}\left(x_{j_i}^T \tilde{\beta}^i - l_{j_i}^*(w_{j_i}^i) \right) e_{j_i}$

Update dual variable: Compute
$w^{i+1} = \arg\min_w \left\{ \left\langle -\eta_i \tilde{g}^i, w - w^i \right\rangle + D_h(w, w^i) \right\}$

(Optional Accounting:) $\bar{\beta}^i \leftarrow (1-\alpha_i)\bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

Randomized Dual Coordinate Mirror Descent

Randomized Dual Coordinate Mirror Descent (RDCMD)

Define the prox function $h(w) := \frac{1}{n}\sum_{i=1}^{n} l_i^*(w_i)$. Initialize with $w^0 = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} l_i^*(w_i)$ and step-size sequences $\{\alpha_i\} \in (0,1]$ and $\{\eta_i\} \in (0,1]$. (Optional: set $\bar{\beta}^{-1} = 0$.)

For iterations $i = 0, 1, \ldots$

Compute Randomized Coordinate of Subgradient of $D(\cdot)$ at w^i

Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(\frac{1}{n}(w^i)^T X\beta + R(\beta) \right) \right\}$

Choose random index. Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Compute subgradient coordinate vector: $\tilde{g}^i \leftarrow \frac{1}{n}\left(x_{j_i}^T \tilde{\beta}^i - \dot{l}_{j_i}^*(w_{j_i}^i) \right) e_{j_i}$

Update dual variable: Compute $w^{i+1} = \arg\min_w \left\{ \left\langle -\eta_i \tilde{g}^i, w - w^i \right\rangle + D_h(w, w^i) \right\}$

(Optional Accounting:) $\bar{\beta}^i \leftarrow (1-\alpha_i)\bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

Recall the Bregman Distance

$$D_h(w, w^i) := h(w) - h(w^i) - \left\langle \nabla h(w^i), w - w^i \right\rangle$$

Randomized Dual Coordinate Mirror Descent

Randomized Dual Coordinate Mirror Descent (RDCMD)

Define the prox function $h(w) := \frac{1}{n}\sum_{i=1}^n l_i^*(w_i)$. Initialize with $w^0 = \arg\min_w \frac{1}{n}\sum_{i=1}^n l_i^*(w_i)$ and step-size sequences $\{\alpha_i\} \in (0,1]$ and $\{\eta_i\} \in (0,1]$. (Optional: set $\bar{\beta}^{-1} = 0$.)

For iterations $i = 0, 1, \ldots$

Compute Randomized Coordinate of Subgradient of $D(\cdot)$ at w^i
Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{\left(\frac{1}{n}(w^i)^T X\beta + R(\beta)\right)\right\}$

Choose random index. Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Compute subgradient coordinate vector: $\tilde{g}^i \leftarrow \frac{1}{n}\left(x_{j_i}^T\tilde{\beta}^i - \dot{l}_{j_i}^*(w_{j_i}^i)\right) e_{j_i}$

Update dual variable: Compute
$w^{i+1} = \arg\min_w \left\{\left\langle -\eta_i \tilde{g}^i, w - w^i\right\rangle + D_h(w, w^i)\right\}$

(Optional Accounting:) $\bar{\beta}^i \leftarrow (1-\alpha_i)\bar{\beta}^{i-1} + \alpha_i\tilde{\beta}^i$.

Randomized Dual Coordinate Mirror Descent

Remarks

- RDCMD takes place completely in the dual space.
- We also used two step-size sequences:
  - $\{\eta_i\}$ is used in the prox subproblem updates of w^i
  - $\{\alpha_i\}$ is used in the optional accounting to update the $\bar{\beta}^i$ values

Equivalence Lemma

Equivalence Lemma

GSFW and RDCMD are equivalent as follows: the iterate sequence of either algorithm exactly corresponds to an iterate sequences of the other.

- In the deterministic case, [Bach 2015] showed that the Frank-Wolfe method for the primal problem is equivalent to mirror descent algorithm for the dual problem under some assumptions
- This provides a new primal interpretation of a randomized dual coordinate descent type of algorithm first introduced in [Shalev-Shwartz, Zhang 2013].

Computational Guarantees

Computational Guarantees

First, Some New Metrics

- Let

$$M := \max_{\beta \in \mathrm{dom} R(\cdot)} \max_{j=1,\ldots,n} \{|x_j^T \beta|\} ,$$

then $M < +\infty$ if $\mathrm{dom} R(\cdot)$ is bounded. Moreover, when $\|x_j\|$ is bounded for any j, M is independent of n.

- Let $\mathcal{W} \subset \mathbb{R}^n$ be the set of "optimal w responses" to values $\beta \in \mathrm{dom} R(\cdot)$ in the saddle-function $\phi(\beta, w)$, namely:

$$\mathcal{W} := \{\hat{w} \in \mathbb{R}^n : \hat{w} \in \arg\max_w \phi(\hat{\beta}, w) \text{ for some } \hat{\beta} \in \mathrm{dom} R(\cdot)\} .$$

- Let $D_{\max}$ be any upper bound on $D_h(\hat{w}, w^0)$ as $\hat{w}$ ranges over all values in $\mathcal{W}$:

$$D_h(\hat{w}, w^0) \leq D_{\max} \quad \text{for all } \hat{w} \in \mathcal{W} .$$

An Upper Bound on $D_{\max}$

Proposition: Upper bound on $D_{\max}$

It holds that

$$D_{\max} \leq \gamma M^2 \ .$$

- However, a much smaller value of $D_{\max}$ can often be easily derived based on the structure of $l_j(\cdot)$. For example, in logistic regression we have simply that $D_{\max} = \ln(2)$.

Convergence Guarantees when $R(\cdot)$ is not Strongly Convex

Theorem: Convergence Guarantees when $R(\cdot)$ is not Strongly Convex

Consider SGFW (or RDCMD) with step-size sequences $\alpha_i = \frac{2(2n+i)}{(i+1)(4n+i)}$ and $\eta_i = \frac{2n}{2n+i+1}$ for $i = 0, 1, \ldots$. Denote

$$\bar{w}^k = \frac{2}{(4n+k)(k+1)} \sum_{i=0}^{k} (2n+i) w^i \ .$$

It holds for all $k \geq 0$ that

$$\mathbb{E}\left[P(\bar{\beta}^k) - D(\bar{w}^k)\right] \leq \frac{8n\gamma M^2}{(4n+k)} + \frac{2n(2n-1)D_{\max}}{(4n+k)(k+1)} \ .$$

Convergence Guarantees when $R(\cdot)$ is not Strongly Convex

Theorem: Convergence Guarantees when $R(\cdot)$ is not Strongly Convex

Consider SGFW (or RDCMD) with step-size sequences $\alpha_i = \frac{2(2n+i)}{(i+1)(4n+i)}$ and $\eta_i = \frac{2n}{2n+i+1}$ for $i = 0, 1, \ldots$. Denote

$$\bar{w}^k = \frac{2}{(4n+k)(k+1)} \sum_{i=0}^{k} (2n+i) w^i \ .$$

It holds for all $k \geq 0$ that

$$\mathbb{E}\left[P(\bar{\beta}^k) - D(\bar{w}^k)\right] \leq \frac{8n\gamma M^2}{(4n+k)} + \frac{2n(2n-1)D_{\max}}{(4n+k)(k+1)} \ .$$

We prove this theorem through the dual lens.

Randomized Dual Coordinate Mirror Descent

Randomized Dual Coordinate Mirror Descent (RDCMD)

Define the prox function $h(w) := \frac{1}{n}\sum_{i=1}^{n} l_i^*(w_i)$. Initialize with $w^0 = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} l_i^*(w_i)$ and step-size sequences $\{\alpha_i\} \in (0,1]$ and $\{\eta_i\} \in (0,1]$. (Optional: set $\bar{\beta}^{-1} = 0$.)

For iterations $i = 0, 1, \ldots$

Compute Randomized Coordinate of Subgradient of $D(\cdot)$ at w^i
Compute $\tilde{\beta}^i \in \arg\min_\beta \left\{ \left(\frac{1}{n}(w^i)^T X\beta + R(\beta) \right) \right\}$

Choose random index. Choose $j_i \in \mathcal{U}[1, \ldots, n]$

Compute subgradient coordinate vector: $\tilde{g}^i \leftarrow \frac{1}{n}\left(x_{j_i}^T \tilde{\beta}^i - \dot{l}_{j_i}^*(w_{j_i}^i) \right) e_{j_i}$

Update dual variable: Compute
$w^{i+1} = \arg\min_w \left\{ \left\langle -\eta_i \tilde{g}^i, w - w^i \right\rangle + D_h(w, w^i) \right\}$

(Optional Accounting:) $\bar{\beta}^i \leftarrow (1-\alpha_i)\bar{\beta}^{i-1} + \alpha_i \tilde{\beta}^i$.

Proof Technique: First-Order Methods (FOM) Naturally Reduce the Primal-Dual Gap Bound

- Previous work on dual coordinate methods need extra assumptions (such as $R(\cdot)$ is strongly convex) and extra mechanics to obtain primal certificates.

- However, first-order methods (stochastic or deterministic, accelerated or non-accelerated, mirror descent or dual averaging) should naturally reduce the primal-dual gap bound, and it is a matter of seeing where this is manifest.

Proof Technique: First-Order Methods (FOM) Naturally Minimize the Primal-Dual Gap Bound, continued

- In standard proof for FOM, one always ends up with

$$D(w) - D(\bar{w}^k) \le \sum_{i=0}^{k} \gamma_i (D(w) - D(w^i)) \le \sum_{i=0}^{k} \gamma_i \langle g^i, w - w^i \rangle \le \cdots \ .$$

- Actually we have

$$\sum_{i=0}^{k} \gamma_i \langle g^i, w - w^i \rangle = \sum_{i=0}^{k} \gamma_i \langle \nabla_w \phi(\tilde{\beta}^i, w^i), w - w^i \rangle$$
$$\ge \sum_{i=0}^{k} \gamma_i \left(\phi(\tilde{\beta}^i, w) - D(w^i) \right) \ge \phi(\bar{\beta}^k, w) - D(\bar{w}^k) \ ,$$

- Choosing $w = \arg\min_w \phi(\bar{\beta}^k, w)$, the right-hand-side becomes $P(\bar{\beta}^k) - D(\bar{w}^k)$.

Proof Technique: Randomized Coordinate Mirror Descent for Non-smooth Function

- There are many results on randomized coordinate descent types of methods for smooth optimization, but not for non-smooth optimization due to the lack of smoothness (used to upper-bound the function).

- One can think of a randomized coordinate of a subgradient as an unbiased estimator of an exact subgradient (up to a scalar multiple). Recall that

$$\tilde{g}^i \leftarrow \frac{1}{n}\left(x_{j_i}^T \tilde{\beta}^i - l_{j_i}^*(w_{j_i}^i)\right) e_{j_i} \ ,$$

whereby

$$n \cdot \mathbb{E}[\tilde{g}^i] = g^i \in \partial D(w^i) \ .$$

- We use the new analysis for stochastic mirror descent algorithm for non-smooth optimization in [Lu 2017].

Convergence Guarantees when $R(\cdot)$ is not Strongly Convex

Theorem: Convergence Guarantees when $R(\cdot)$ is not Strongly Convex

Consider SGFW (or RDCMD) with step-size sequences $\alpha_i = \frac{2(2n+i)}{(i+1)(4n+i)}$ and $\eta_i = \frac{2n}{2n+i+1}$ for $i = 0, 1, \ldots$. Denote

$$\bar{w}^k = \frac{2}{(4n+k)(k+1)} \sum_{i=0}^{k} (2n+i) w^i \ .$$

It holds for all $k \geq 0$ that

$$\mathbb{E}\left[P(\bar{\beta}^k) - D(\bar{w}^k)\right] \leq \frac{8n\gamma M^2}{(4n+k)} + \frac{2n(2n-1)D_{\max}}{(4n+k)(k+1)} \ .$$

- We prove the theorem through the dual lens.

Relative Strong Convexity

Definition: Relative Strong Convexity [Lu, Freund, Nesterov 2018]

$f(\cdot)$ is μ-strongly convex relative to $h(\cdot)$ if for any x, y, there is a scalar μ for which

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \mu D_h(y, x).$$

- This is a stronger definition than $h(\cdot)$ is strongly convex with respect to a norm and $f(\cdot)$ is strongly convex with respect to that norm.
- But it is only with this stronger definition that we have a linear convergence result for the mirror descent algorithm ([Lu, Freund, Nesterov 2018]), but see also [Hanzely and Richtarik 2018].

Coordinate-Wise Relative Smoothness

Definition: Coordinate-Wise Relative Smoothness (Adapted from [Hanzely and Richtarik 2018])

$f(\cdot)$ is coordinate-wise σ-smooth relative to a separable convex reference function $h(\cdot)$ if there is a scalar σ such that for any x, scalar t and coordinate j and $y = x + te_j$ we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \sigma D_h(y, x).$$

Convergence Guarantees when $R(\cdot)$ is Strongly Convex

Theorem: Convergence Guarantees when $R(\cdot)$ is Strongly Convex

Assume $D(w)$ is σ coordinate-wise smooth relative to $h(w)$. Consider the Randomized Dual Coordinate Mirror Descent method with step-size $\eta_i = \frac{1}{\sigma}$ and $\alpha_i = \frac{\sigma^i}{\sigma^{i+1} - (\sigma - 1/n)^{i+1}}$. Denote

$$\bar{w}^k \leftarrow \frac{1}{\sum_{i=0}^k \left(\frac{n\sigma}{n\sigma - 1}\right)^i} \sum_{i=0}^k \left(\frac{n\sigma}{n\sigma - 1}\right)^i w^i \ ,$$

then we have

$$\mathbb{E}\left[P(\bar{\beta}^k) - D(\bar{w}^k)\right] \le \frac{D_{\max}}{\left(1 + \frac{1}{n\sigma - 1}\right)^k - 1} \le \frac{\gamma M^2}{\left(1 + \frac{1}{n\sigma - 1}\right)^k - 1} .$$

A simpler (but looser) bound is simply

$$\frac{D_h(x, x^0)}{\left(1 + \frac{1}{n\sigma - 1}\right)^k - 1} \le n\sigma \left(1 - \frac{1}{n\sigma}\right)^k D_h(x, x^0) \ .$$

Convergence Guarantees when $R(\cdot)$ is Strongly Convex

Corollary

(1) If $R(\cdot)$ is not separable, let $\sigma = \frac{\lambda_{\max}(XX^T)}{n\mu\gamma} + 1$, then the Theorem implies

$$\mathbb{E}\left[P(\bar{\beta}^k) - D(\bar{w}^k)\right] \leq \frac{M^2\lambda_{\max}(XX^T)}{\mu}\left(1 - \frac{\lambda_{\max}(XX^T)}{\mu\gamma}\right)^k .$$

(2) If $R(\cdot)$ is separable, let $\sigma = \frac{\max_j \|X_j\|_2^2}{n\mu\gamma} + 1$, then the Theorem implies

$$\mathbb{E}\left[P(\bar{\beta}^k) - D(\bar{w}^k)\right] \leq \frac{M^2 \max_j \|X_j\|_2^2}{\mu}\left(1 - \frac{\max_j \|X_j\|_2^2}{\mu\gamma}\right)^k .$$

Some Discussions/Extensions

- Both the algorithm and the analysis can be easily extended to the mini-batch setting.
- We can also generalize the algorithm and analysis to non-uniform sampling.
- When $R(\cdot)$ is strongly convex, we can also achieve accelerated linear convergence by utilizing the technique developed in [Lin, Lu, Xiao 2015].
- The unaccelerated version of [Lin, Lu, Xiao 2015] can be viewed as randomized dual coordinate mirror descent with the reference function $h(w) = \frac{1}{n}\sum_{j=1}^{n} l_j^*(w_j) + \frac{\lambda}{2}\|w\|^2$ for some λ, while we here use randomized dual coordinate mirror descent with reference function $h(w) = \frac{1}{n}\sum_{j=1}^{n} l_j^*(w_j)$.

Contribution/Summary

Contribution/Summary:

- Stochastic Generalized Frank-Wolfe Method with Substitute Gradient
- Randomized Dual Coordinate Mirror Descent Algorithm
- Equivalence of SGFW and RDCMD, which leads to new primal interpretations of dual coordinate methods
- $O(\frac{1}{\varepsilon})$ Stochastic Frank-Wolfe Method
- Linear convergence result when $R(\cdot)$ is strongly convex
- We show that these FOMs inherently reduce the primal-dual gap bound
- Computational guarantees for randomized coordinate descent for minimizing non-smooth functions

References

- Francis Bach, *Duality between subgradient and conditional gradient methods*
- Filip Hanzely and Peter Richtárik, *Fastest rates for stochastic mirror descent methods*
- Elad Hazan and Satyen Kale, *Projection-free online learning*
- Elad Hazan and Haipeng Luo, *Variance-reduced and projection-free stochastic optimization*
- Martin Jaggi, *Revisiting Frank-Wolfe: Projection-free sparse convex optimization*
- Guanghui Lan and Yi Zhou, *Conditional gradient sliding for convex optimization*
- Qihang Lin, Zhaosong Lu, and Lin Xiao, *An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization*
- Haihao Lu, *" relative-continuity" for non-lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent*
- Haihao Lu, Robert M Freund, and Yurii Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*
- Zhaosong Lu and Lin Xiao, *On the complexity analysis of randomized block-coordinate descent methods*
- Yu Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*
- Peter Richtarik and Martin Takac, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*
- Shai Shalev-Shwartz and Tong Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*
- Shai Shalev-Shwartz and Tong Zhang, *Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization*