Faster optimization through adaptive importance sampling [Perekrestenko et al., 2017]

Student:

Dmytro Perekrestenko

Supervisors:

Prof. Martin Jaggi
Prof. Volkan Cevher

May 25, 2018

Outline

Primal-dual setting [Dünner et al., 2016]

The following pair of dual to each other optimization problems is considered:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left[D(\boldsymbol{\alpha}) := f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i) \right],$$

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \left[P(\boldsymbol{w}) := f^*(\boldsymbol{w}) + \sum_i g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) \right].$$

Examples:

- SVM: $f^*(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$, $g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) = \max(0, 1 - y_i \boldsymbol{a}_i^\top \boldsymbol{w})$.
- LASSO: $f(A\boldsymbol{\alpha}) = \|\boldsymbol{y} - A\boldsymbol{\alpha}\|_2^2$, $g_i(\alpha_i) = \lambda|\alpha_i|$.

Goal

Find a ϵ_P-suboptimal $\boldsymbol{w}$ or ϵ_D-suboptimal $\boldsymbol{\alpha}$, i.e., $P(\boldsymbol{w}) - P(\boldsymbol{w}^\star) \leq \epsilon_P$ or $D(\boldsymbol{\alpha}) - D(\boldsymbol{\alpha}^\star) \leq \epsilon_D$.

Optimality conditions and classic algorithm

Optimality Conditions

Under assumption of strong duality: $P(\boldsymbol{w}^\star) = -D(\boldsymbol{\alpha}^\star)$ and

$$\boldsymbol{w}^\star \in \partial f(A\boldsymbol{\alpha}^\star) \qquad \alpha_i^\star \in \partial g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}^\star) \text{ for all } i \in [n]$$

Duality gap is defined as: $G(\boldsymbol{\alpha}, \boldsymbol{w}) := P(\boldsymbol{w}(\boldsymbol{\alpha})) - (-D(\boldsymbol{\alpha}))$.
It can act as a stopping criterion $G(\boldsymbol{\alpha}) \geq P(\boldsymbol{w}(\boldsymbol{\alpha})) - P(\boldsymbol{w}^\star)$.

Algorithm 1 Stochastic Coordinate Descent

1: let $\boldsymbol{\alpha}^{(0)} = 0 \in \mathbb{R}^n$, $\boldsymbol{w}^{(0)} = \boldsymbol{w}(\boldsymbol{\alpha}^{(0)})$
2: **for** $t = 0, 1, \ldots T$ **do**
3: Sample $i \in [n]$ randomly according to distribution $\boldsymbol{p}$
4: Find $\Delta\alpha_i$ minimizing $D(\boldsymbol{\alpha}^{(t)} + \boldsymbol{e}_i \Delta\alpha_i)$
5: $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \boldsymbol{e}_i \Delta\alpha_i$, $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}(\boldsymbol{\alpha}^{(t+1)})$
6: **end for**

Motivation

The Limitations of Existing Algorithms

- sampling of the active datapoint uniformly at random in each iteration
- the convergence rate is negatively affected

Stochastic Optimization with Adaptive Importance Sampling

- adaptively changes the sampling probability distribution according to the data and values of the dual variables
- improved practical and theoretical convergence rates

Basic definitions

Definition (Dual Residual, ([Csiba et al., 2015], Def. 1))

Consider our primal-dual setting. Given dual variable α, the i-th dual residue on iteration t is given by:

$$\kappa_i^{(t)} = u_i^{(t)} - \alpha_i^{(t)},$$

where $u_i^{(t)} = \nabla g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)})$.

Definition (Coherent probability vector, ([Csiba et al., 2015], Def. 2))

We say that probability vector $\boldsymbol{p}^{(t)} \in \mathbb{R}^n$ is coherent with the dual residue vector $\boldsymbol{\kappa}^{(t)}$ if for all $i \in [n]$, we have $\kappa_i^{(t)} \neq 0 \quad \rightarrow \quad p_i^{(t)} > 0$.

Definition (t-support set)

We call set I_t: $I_t = \{i \in [n] : \kappa_i^{(t)} \neq 0\} \subseteq [n]$ a t-support set.

Outline

Core lemma

Lemma (A generalization of ([Csiba et al., 2015], Lemma 3))

Consider Stochastic Coordinate Descent. Let f be $1/\beta$-smooth and g_i be μ_i-strongly convex with convexity parameter $\mu_i \geq 0\ \forall i \in [n]$. For the case $\mu_i = 0$ we require g_i to have a bounded support. Then for any iteration t, any sampling distribution $\boldsymbol{p}^{(t)}$ coherent with $\boldsymbol{\kappa}^{(t)}$ and any $\theta \in [0, \min_{i \in I_t} p_i^{(t)}]$ it holds that:

$$\mathbb{E}[D(\boldsymbol{\alpha}^{(t+1)})|\boldsymbol{\alpha}^{(t)}] \leq D(\boldsymbol{\alpha}^{(t)}) - \theta G(\boldsymbol{\alpha}^{(t)}) + \frac{\theta^2 n^2}{2} F^{(t)},$$

where

$$F^{(t)} = \frac{1}{n^2 \beta \theta} \sum_{i \in I_t} \left(\frac{\theta(\mu_i \beta + \|\boldsymbol{a}_i\|^2)}{p_i^{(t)}} - \mu_i \beta \right) |\kappa_i^{(t)}|^2.$$

Outline

General convex objectives

Theorem

Consider Stochastic Coordinate Descent. Assume f is $\frac{1}{\beta}$-smooth function. Then, if g_i^* is L_i-Lipschitz for each i and $\boldsymbol{p}^{(t)}$ is coherent with $\boldsymbol{\kappa}^{(t)}$, it suffices to have a total number of iterations of

$$T \geq \max\left\{0, \frac{1}{p_{\min}} \log\left(\frac{2\epsilon_D^0}{n^2 p_{\min} F^\circ}\right)\right\} + \frac{5F^\circ n^2}{\epsilon} - \frac{1}{p_{\min}}$$

or alternatively

$$T \geq \frac{5F^\circ n^2}{\epsilon} + \frac{5\epsilon_D^{(0)}}{\epsilon p_{\min}} - \frac{1}{p_{\min}}$$

to obtain a duality gap $G(\bar{\alpha}) \leq \epsilon$, where ϵ_D^0 is the initial dual suboptimality and F° is an upper bound on $\mathbb{E}[F^{(t)}]$ taken over all coordinates at $1, \ldots, T$ algorithm iterations.

General convex objectives

From this theorem we can recover as a special case:

1. ([Dünner et al., 2016], Theorem 9) by setting $p_i = \frac{1}{n}$ and using R instead of $\|\boldsymbol{a}_i\|$ ($R \geq \|\boldsymbol{a}_i\| \; \forall i \in [n]$).

2. ([Shalev-Shwartz and Zhang, 2013], Theorem 2) is a special case of ([Dünner et al., 2016], Theorem 9) for quadratic regularizer.

3. ([Zhao and Zhang, 2014], Theorem 5) by setting $p_i = \frac{L_i}{\sum_j L_j}$.

Maximizing rate in case of infinitesimal ϵ

Recall that the number of iterations sufficient to achieve ϵ-accuracy is:

$$T \geq \max\left\{0, \frac{1}{p_{\min}} \log\left(\frac{2\epsilon_D^0}{n^2 p_{\min} F^\circ}\right)\right\} + \frac{5F^\circ n^2}{\epsilon} - \frac{1}{p_{\min}}$$

In limit $\epsilon \to 0$, the only significant term is $\frac{5F^\circ n^2}{\epsilon}$, therefore we want to minimize $F^{(t)}$, which consequently minimizes F°. We solve the following optimization problem:

$$\boldsymbol{p}^{(t)} = \arg\min_{\boldsymbol{p}^{(t)}} F^{(t)} := \arg\min_{\boldsymbol{p}^{(t)}} \frac{1}{n^2\beta} \sum_i \left(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^t}\right).$$

Optimal sampling distributions

1) For uniform sampling $\forall t$:

$$p_i := \frac{1}{n}; \quad F^{\circ}_{\text{unif}} = \mathbb{E}\left[\frac{1}{\beta}\sum_i \left(\frac{|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|^2}{n}\right)\right];$$

2) For importance sampling $\forall t$:

$$p_i := \frac{L_i\|\boldsymbol{a}_i\|}{\sum_j L_j\|\boldsymbol{a}_j\|}; \quad F^{\circ}_{\text{imp}} = \mathbb{E}\left[\frac{1}{\beta}\sum_i \left(\frac{|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|}{L_i n}\right)\sum_j \left(\frac{L_j\|\boldsymbol{a}_j\|}{n}\right)\right];$$

3) For adaptive sampling $\forall t$:

$$p_i^{(t)} := \frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}; \quad F^{\circ}_{\text{ada}} = \mathbb{E}\left[\frac{1}{\beta}\left(\sum_i \frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{n}\right)^2\right];$$

Support set uniform sampling

Let's assume that the size of the t-support set never exceeds some $m \in [1, n]$ and compare two sampling methods:

- Uniform sampling: $p_i^{(t)} = \frac{1}{n}$; $p_{\min} = \frac{1}{n}$,

  $$F^{(t)} = \frac{1}{n^2\beta} \sum_{i \in I_t} \left(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^{(t)}} \right) = \frac{1}{n\beta} \sum_{i \in I_t} |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2 \leq F_{\text{unif}}$$

  $$\text{Number of iterations: } T \geq \max \left\{ 0, n \log \left(\frac{2\epsilon_D^0}{nF_{\text{unif}}} \right) \right\} + \frac{5n^2 F_{\text{unif}}}{\epsilon}$$

- Support set uniform: $\begin{cases} p_i^{(t)} = \frac{1}{m}, & \text{if } \kappa_i^{(t)} \neq 0 \\ p_i^{(t)} = 0, & \text{otherwise} \end{cases}$; $p_{\min} = \frac{1}{m}$,

  $$F^{(t)} = \frac{1}{n^2\beta} \sum_{i \in I_t} \left(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^{(t)}} \right) = \frac{m}{n^2\beta} \sum_{i \in I_t} |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2 \leq \frac{m}{n} F_{\text{unif}}$$

  $$\text{Number of iterations: } T \geq \max \left\{ 0, m \log \left(\frac{2\epsilon_D^{(0)}}{nF_{\text{unif}}} \right) \right\} + \frac{5nmF_{\text{unif}}}{\epsilon}$$

Maximizing rate in case of constant ϵ

The number of iterations T is directly proportional to F° and $1/p_{\min}$, the optimal distribution $\boldsymbol{p}$ should minimize F° and $1/p_{\min}$ at the same time. We define mixed distribution as:

$$\begin{cases} p_i^{(t)} = \frac{\sigma}{m} + (1-\sigma)\frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}, & \text{if } \kappa_i^{(t)} \neq 0 \\ p_i^{(t)} = 0, & \text{otherwise} \end{cases}$$

where $\sigma \in [0, 1]$. This distribution gives us the following bounds on F° and $1/p_{\min}$:

$$F^\circ_{\text{mix}} \leq \frac{F^\circ_{\text{ada}}}{1-\sigma} \qquad \frac{1}{p_{\min}} \leq \frac{m}{\sigma},$$

and bound on the number of iterations:

$$T \geq \frac{5F^\circ_{\text{ada}} n^2}{\epsilon(1-\sigma)} + \frac{\epsilon_D^{(0)} m}{\epsilon\sigma}.$$

Outline

Gap-wise sampling

Remark

The duality gap can be decomposed as a sum of coordinate-wise duality gaps:

$$G(\alpha) = \sum_i G_i(\alpha_i, \boldsymbol{w}) = \sum_i \left(g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) + g_i(\alpha_i) + \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w} \right)$$

Definition (Nonuniformity measure, [Osokin et al., 2016])

The nonuniformity measure $\chi(\boldsymbol{x})$ of a vector $\boldsymbol{x} \in \mathbb{R}^n$, is defined as:

$$\chi(\boldsymbol{x}) := \sqrt{1 + n^2 \mathrm{Var}[\boldsymbol{p}]},$$

where $\boldsymbol{p} := \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}$ is the probability vector obtained by normalizing $\boldsymbol{x}$.

Gap-wise sampling

Theorem

Consider Stochastic Coordinate Descent. Assume f is $\frac{1}{\beta}$-smooth function. Then, if g_i^* is L_i-Lipschitz for each i and $p_i^{(t)} := \frac{G_i(\boldsymbol{\alpha}^{(t)})}{G(\boldsymbol{\alpha}^{(t)})}$ then on each iteration it holds that

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \frac{C + 2n\epsilon_D^{(0)}}{t + 2n},$$

where C is an upper bound on $\mathbb{E}\left[\frac{2n\chi(\vec{\boldsymbol{F}})\sum_i \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2}{(\chi(\vec{\boldsymbol{G}}))^3\beta}\right]$, where the expectation is taken over the random choice of the sampled coordinate at iterations $1, \ldots, t$ of the algorithm. Here $\vec{\boldsymbol{G}}$ and $\vec{\boldsymbol{F}}$ are defined as:

$$\vec{\boldsymbol{G}} := (G_i(\boldsymbol{\alpha}^{(t)}))_{i=1}^n, \qquad \vec{\boldsymbol{F}} := (\|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2)_{i=1}^n.$$

Gap-wise vs uniform

From intermediate result in the proof of Theorem 1, uniform distribution ($p_i = 1/n$) has the following convergence rate:

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \frac{\frac{2n}{\beta}\mathbb{E}\left[\sum_i |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2\right] + 2n\epsilon_D^{(0)}}{2n+t}.$$

The rate of gap-wise sampling depends on non-uniformity measures $\chi(\vec{\boldsymbol{G}})$ and $\chi(\vec{\boldsymbol{F}})$:

$$\mathbb{E}[\epsilon_D^{t}] \leq \frac{\frac{2n}{\beta}\mathbb{E}\left[\frac{\chi(\vec{\boldsymbol{F}})}{(\chi(\vec{\boldsymbol{G}}))^3}\sum_i |\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2\right] + 2n\epsilon_D^{(0)}}{2n+t}.$$

In the worst case scenario when variance is maximal in $(|\kappa_i^{(t)}|^2\|\boldsymbol{a}_i\|^2)_{i=1}^n$, $\chi(\vec{\boldsymbol{F}}) \approx \sqrt{n}$, the rate of gap-wise sampling is better than of uniform only when the gaps are non-uniform enough i.e., $\chi(\vec{\boldsymbol{G}}) \geq n^{\frac{1}{6}}$.

Outline

Lasso problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\Big[D(\boldsymbol{\alpha}) := f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i)\Big],$$

$$\min_{\boldsymbol{w}\in\mathbb{R}^d}\Big[P(\boldsymbol{w}) := f^*(\boldsymbol{w}) + \sum_i g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w})\Big].$$

Here $f(A\boldsymbol{\alpha}) = \frac{1}{2n}\|A\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2$ and $g_i(\alpha_i) = \lambda|\alpha_i|$. To satisfy requirements of the theorem we use "Lipschitzing trick" [Dünner et al., 2016] on $g_i(\alpha_i)$ to make g_i^* a Lipschitz function:

$$\bar{g}_i(\alpha_i) = \begin{cases} \lambda|\alpha_i|, & \text{if } |\alpha_i| \le B \\ +\infty, & \text{otherwise} \end{cases}$$

The $\bar{g}_i$-conjugate will be:

$$\bar{g}_i^*(u_i) = \max_{\alpha_i:|\alpha_i|\le B} u_i\alpha_i - \lambda|\alpha_i| = B\big[|u_i| - \lambda\big]_+$$

Stochastic Coordinate Descent - fixed

Algorithm 2 Stochastic Coordinate Descent

let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
for t = 0,1,... **do**
    sample j from $[d]$ according to distribution $\boldsymbol{p}$
    let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
    $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
    $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
end for

- **uniform** $p_i = \frac{1}{d}$
- **importance** $p_i = \frac{L_i \|\boldsymbol{a}_i\|}{\sum_j L_j \|\boldsymbol{a}_j\|}$
- (*heuristic*) **gap-init** $p_i = \frac{G_i^{(0)}}{\sum_j G_j^{(0)}}$

Stochastic Coordinate Descent - adaptive

Algorithm 3 Stochastic Coordinate Descent (adaptive)

1: let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
2: **for** t = 0,1,... **do**
3: calculate absolute values of dual residuals $|\kappa_j^{(t)}|$ for all $j \in [d]$
4: generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

$$p_i^{(t)} = \frac{|\kappa_i^{(t)}| \|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}| \|\boldsymbol{a}_j\|}$$

5: sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
6: let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
7: $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
8: $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
9: **end for**

Stochastic Coordinate Descent - adaptive

Algorithm 4 Stochastic Coordinate Descent (supportSet-uniform)

1: let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
2: **for** t = 0,1,... **do**
3: calculate absolute values of dual residuals $|\kappa_j^{(t)}|$ for all $j \in [d]$
4: find t-support set $I_t = \{i \in [d] : \kappa_i^{(t)} \neq 0\} \subseteq [d]$
5: generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

$$\begin{cases} p_i^{(t)} = \frac{1}{|I_t|}, & \text{if } \kappa_i^{(t)} \neq 0 \\ p_i^{(t)} = 0, & \text{otherwise} \end{cases}$$

6: sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
7: let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
8: $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$, $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
9: **end for**

Stochastic Coordinate Descent - adaptive

Algorithm 5 Stochastic Coordinate Descent (ada-uniform)

1: let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
2: **for** t = 0,1,... **do**
3: calculate absolute values of dual residuals $|\kappa_j^{(t)}|$ for all $j \in [d]$
4: find t-support set $I_t = \{i \in [d] : \kappa_i^{(t)} \neq 0\} \subseteq [d]$
5: generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$:

$$\begin{cases} p_i^{(t)} = \frac{1}{2|I_t|} + \frac{|\kappa_i^{(t)}| \|\boldsymbol{a}_i\|}{2\sum_j |\kappa_j^{(t)}| \|\boldsymbol{a}_j\|}, & \text{if } \kappa_i^{(t)} \neq 0 \\ p_i^{(t)} = 0, & \text{otherwise} \end{cases}$$

6: sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
7: let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
8: $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$, $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
9: **end for**

Stochastic Coordinate Descent - adaptive

Algorithm 6 Stochastic Coordinate Descent (ada-gap)

let $\boldsymbol{\alpha}^{(0)} = 0$, $\boldsymbol{w}^{(0)} = \nabla f(A\boldsymbol{\alpha}^{(0)})$
for t = 0,1,... **do**
    calculate feature-wise duality gaps $G_j^{(t)}$ for all $j \in [d]$
    generate adapted probabilities distribution $\boldsymbol{p}^{(t)}$: $p_i^{(t)} = \frac{G_i^{(t)}}{\sum_j G_j^{(t)}}$
    sample j from $[d]$ according to $\boldsymbol{p}^{(t)}$
    let $z_j = (\nabla f(\boldsymbol{\alpha}^{(t)}))_j$
    $\alpha_j^{(t+1)} = s_\lambda(\alpha_j^{(t)} - z_j)$
    $\boldsymbol{w}^{(t+1)} = \nabla f(A\boldsymbol{\alpha}^{(t+1)})$
end for

Datasets and algorithm complexity

Dataset	Features	Points	nnz/(nd)	mean of $\|\boldsymbol{a}_i\|$	Var($\|\boldsymbol{a}_i\|$)
mushrooms	112	8124	18.8%	31.35	545
rcv1*	809	7438	0.3%	2.58	17.3

Algorithm	**Cost of an Epoch**	**Mode**
Lasso uniform	$O(\text{nnz})$	uniform
Lasso importance	$O(\text{nnz} + n\log(n))$	fixed non-uniform
Lasso gap-init	$O(\text{nnz} + n\log(n))$	fixed non-uniform
Lasso supportSet-uniform	$O(n \cdot \text{nnz})$	adaptive
Lasso adaptive	$O(n \cdot \text{nnz})$	adaptive
Lasso ada-uniform	$O(n \cdot \text{nnz})$	adaptive
Lasso ada-division	$O(\text{nnz} + n\log(n))$	adaptive
Lasso ada-gap	$O(n \cdot \text{nnz})$	adaptive

Numerical experiment - fixed

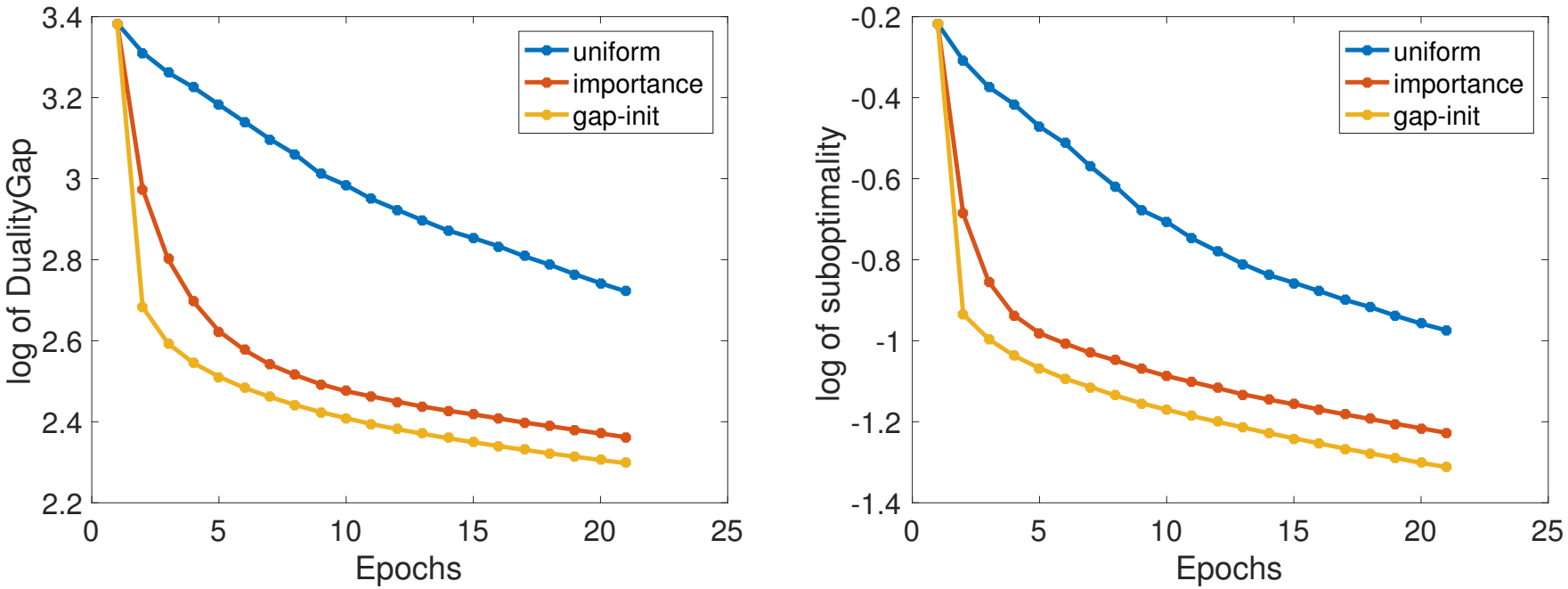

Figure: Lasso. Comparison of different fixed distribution versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - rcv1 dataset

Numerical experiment - fixed

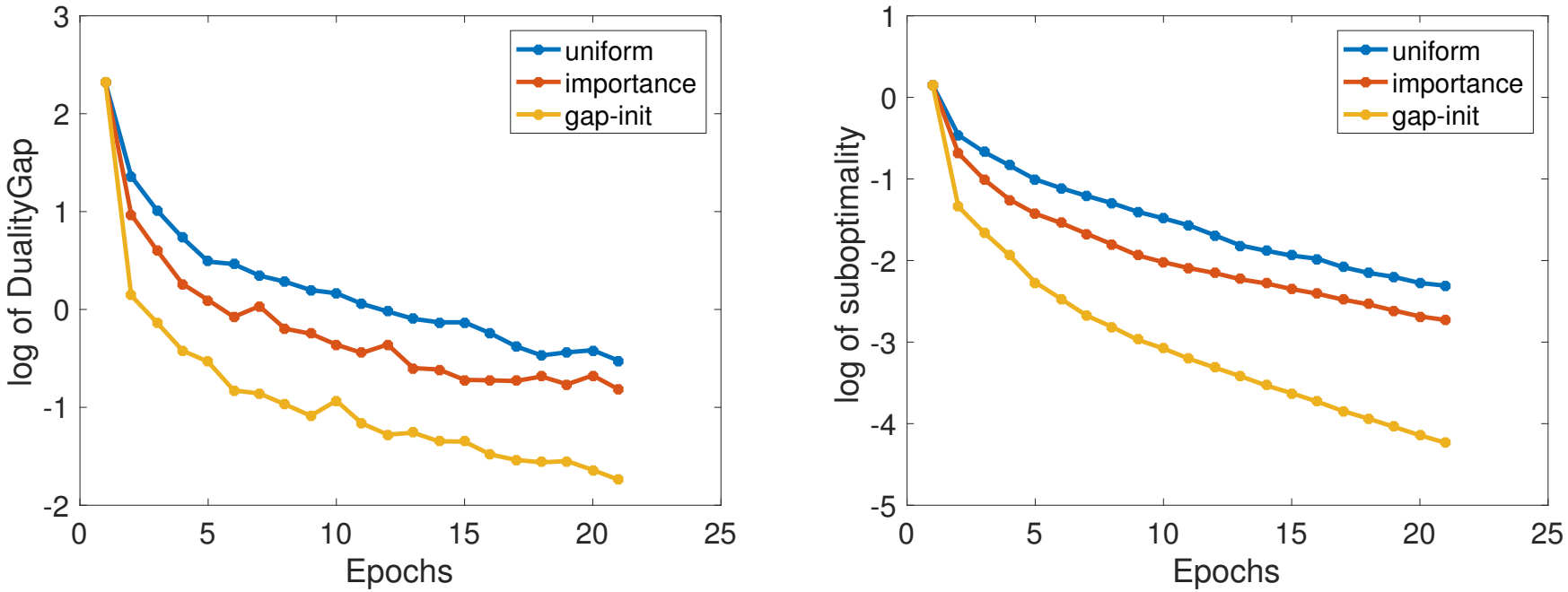

Figure: Lasso. Comparison of different fixed distribution versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - mushrooms dataset

Numerical experiment - adaptive

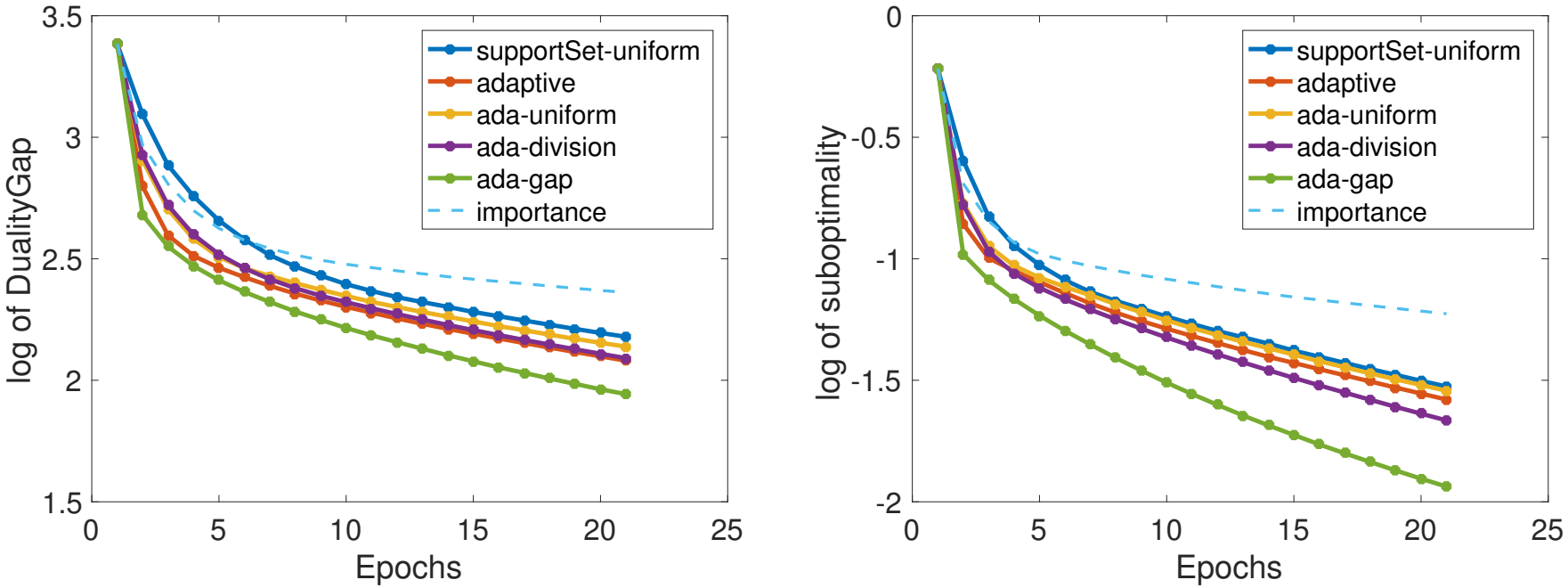

Figure: Lasso. Comparison of different adaptive versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - rcv1 dataset

Numerical experiment - adaptive

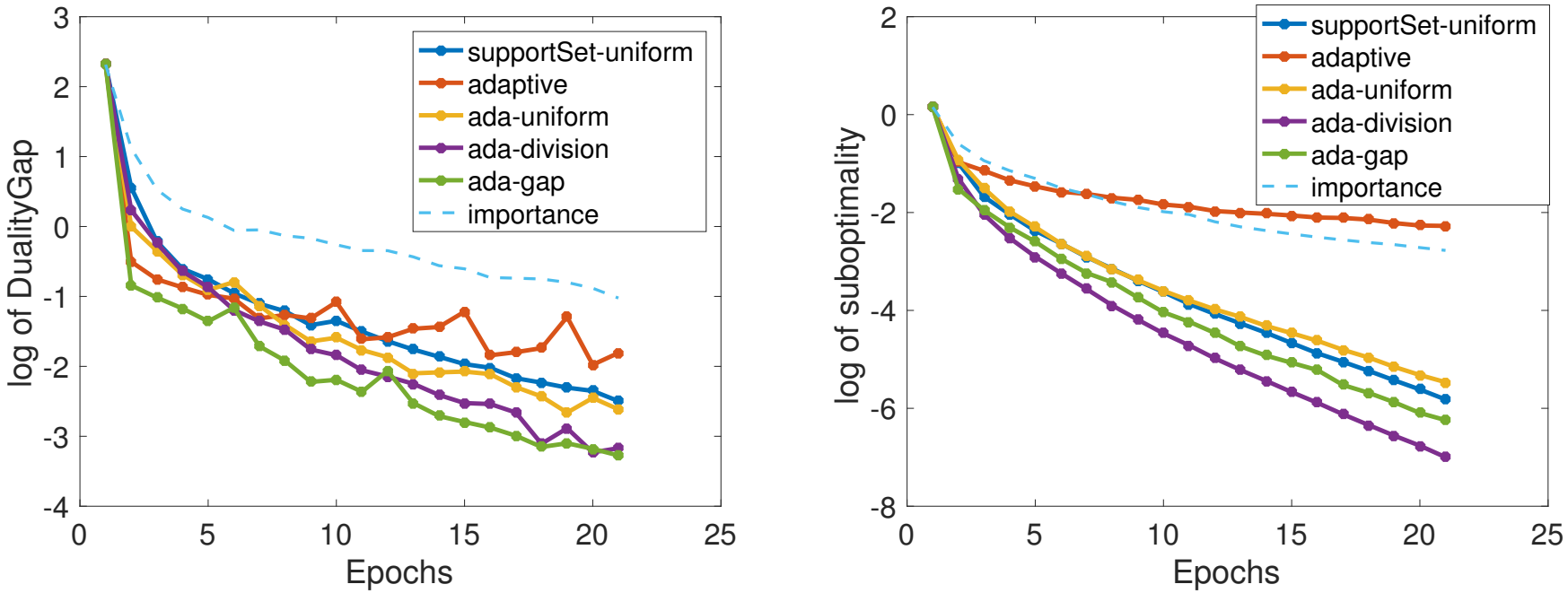

Figure: Lasso. Comparison of different adaptive versions of Stochastic Coordinate Descent Algorithm based on duality gap(left) and suboptimality(right) measure - mushrooms dataset

Outline

Summary and future work

Summary

- studied SCD with adaptive sampling
- proposed new adaptive and fixed nonuniform sampling schemes
- analyzed their theoretical convergence rates
- showed in practice that they outperform the conventional sampling schemes

Future work

- to find a better solution for the constant ϵ optimization problem
- to find a relation between dual residuals sampling and gap-wise sampling
- to apply the theory to hinge loss SVM

References I


Csiba, D., Qu, Z., and Richtárik, P. (2015).
Stochastic Dual Coordinate Ascent with Adaptive Probabilities.
In *ICML 2015 - Proceedings of the 32th International Conference on Machine Learning*.


Dünner, C., Forte, S., Takáč, M., and Jaggi, M. (2016).
Primal-Dual Rates and Certificates.
In *ICML 2016 - Proceedings of the 33th International Conference on Machine Learning*.


Osokin, A., Alayrac, J.-B., Lukasewitz, I., Dokania, P., and Lacoste-Julien, S. (2016).
Minding the Gaps for Block Frank-Wolfe Optimization of Structured SVMs.
In *ICML 2016 - Proceedings of the 33th International Conference on Machine Learning*, pages 593–602.

References II


Perekrestenko, D., Cevher, V., and Jaggi, M. (2017).
Faster coordinate descent via adaptive importance sampling.
Proc. of the 20th International Conference on Artificial Intelligence and Statistics, 54.


Shalev-Shwartz, S. and Zhang, T. (2013).
Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization.
JMLR, 14:567–599.


Zhao, P. and Zhang, T. (2014).
Stochastic Optimization with Importance Sampling.
arXiv.