# Faster Coordinate Descent via Adaptive Importance Sampling

**Dmytro Perekrestenko, Volkan Cevher, Martin Jaggi**

pdmytro@nari.ee.ethz.ch, volkan.cevher@epfl.ch, martin.jaggi@epfl.ch

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

**Abstract**

Coordinate descent methods employ random partial updates of decision variables in order to solve huge-scale convex optimization problems. In this work, we introduce new *adaptive* rules for the random selection of their updates. By adaptive, we mean that our selection rules are based on the dual residual or the primal-dual gap estimates and can change at each iteration. We theoretically characterize the performance of our selection rules and demonstrate improvements over the state-of-the-art, and extend our theory and algorithms to general convex objectives. Numerical evidence with hinge-loss support vector machines and Lasso confirm that the practice follows the theory.

## Problem

We solve the problem of the following template:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i), \tag{1}$$

where $A$ is the data matrix, $f$ is a smooth convex function, and each $g_i$ is a general convex function.

## Primal-dual framework

We are working in the following primal-dual optimization framework ([Dünner et al., 2016]):

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

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \Big[\mathcal{O}_B(\boldsymbol{w}) := f^*(\boldsymbol{w}) + \sum_i g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w})\Big], \tag{B}$$

where we have $A = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_n]$.

## Duality gap

The duality gap is the difference between primal and dual solutions:

$$G(\boldsymbol{\alpha}, \boldsymbol{w}) := \mathcal{O}_A(\boldsymbol{\alpha}) - (-\mathcal{O}_B(\boldsymbol{w})), \tag{2}$$

which can be written as a sum of coordinate-wise gaps:

$$G(\boldsymbol{\alpha}) = \sum_i G_i(\alpha_i) := \sum_i \Big(g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}) + g_i(\alpha_i) + \alpha_i \boldsymbol{a}_i^\top \boldsymbol{w}\Big). \tag{3}$$

## Preliminaries

**Definition** (Dual Residual. A generalization of [Csiba et al., 2015])**.** *$i$-th dual residue on iteration $t$ is given by:*

$$\kappa_i^{(t)} := \min_{u\in\partial g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w}^{(t)})} |u - \alpha_i^{(t)}|.$$

**Definition** (Nonuniformity measure, [Osokin et al., 2016])**.** *The nonuniformity measure of a vector $\boldsymbol{x} \in \mathbb{R}^n$:*

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

**Definition** (Averaged residual)**.** *Averaged residual on iteration $t$ is given by:*

$$F^{(t)} := \frac{1}{n^2\beta} \sum_{i\in I_t} \Big(\frac{|\kappa_i^{(t)}|^2 \|\boldsymbol{a}_i\|^2}{p_i^{(t)}}\Big). \tag{4}$$

**Algorithm 1** Coordinate Descent

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

## Adaptive Sampling residual-based CD

**Theorem.** *If $f$ is a $\frac{1}{\beta}$-smooth and $g_i^*$ is $L_i$-Lipschitz for each $i$, then the residual-based CD iterates satisfy*

$$\mathbb{E}[\varepsilon_A^{(t)}] \le \frac{2F^\circ n^2 + \frac{2\varepsilon_A^{(0)}}{p_{\min}}}{\frac{2}{p_{\min}} + t}. \tag{5}$$

*A* duality gap $G(\bar{\boldsymbol{\alpha}}) \le \varepsilon$ *is reached after an overall number of iterations $T$ whenever*

$$T \ge \max\left\{0, \frac{1}{p_{\min}} \log\Big(\frac{2\varepsilon_A^{(0)}}{n^2 p_{\min} F^\circ}\Big)\right\} + \frac{5F^\circ n^2}{\varepsilon} - \frac{1}{p_{\min}}. \tag{6}$$

*Here $F^\circ$ is an upper bound on $\mathbb{E}[F^{(t)}]$.*

## Adaptive Sampling gap-based CD

**Theorem.** *If $f$ be a $\frac{1}{\beta}$-smooth and $g_i^*$ is $L_i$-Lipschitz for each $i$, then the gap-based CD iterates satisfy*

$$\mathbb{E}[\varepsilon_A^{(t)}] \le \frac{2F_g^\circ n^2 + 2n\varepsilon_A^{(0)}}{t + 2n}, \tag{7}$$

*where $F_g^\circ$ is an upper bound on $\mathbb{E}\Big[F_g^{(t)}\Big]$. The $\overrightarrow{\boldsymbol{G}}$ and $\overrightarrow{\boldsymbol{F}}$ are defined as follows:*

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

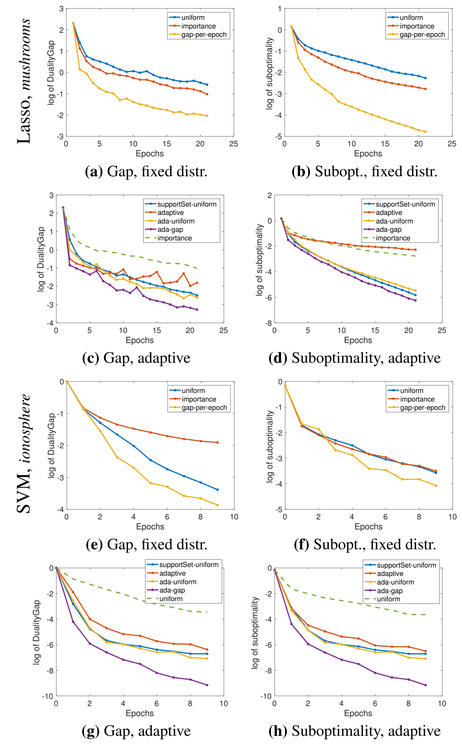
*and $F_g^{(t)}$ is:*

$$F_g^{(t)} := \frac{\chi(\overrightarrow{\boldsymbol{F}})}{n\beta(\chi(\overrightarrow{\boldsymbol{G}}))^3} \sum_i \|\boldsymbol{a}_i\|^2 |\kappa_i^{(t)}|^2. \tag{8}$$

## Variations along the theme

- *uniform* - sample uniformly at random.
- *supportSet uniform* - $p_i^{(t)} := \begin{cases} \frac{1}{m_t}, & \text{if } \kappa_i^{(t)} \neq 0 \\ 0, & \text{otherwise.} \end{cases}$
- *adaptive* - $p_i^{(t)} := \frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}$.
- *ada-uniform* - $p_i^{(t)} := \begin{cases} \frac{0.5}{m_t} + 0.5\frac{|\kappa_i^{(t)}|\|\boldsymbol{a}_i\|}{\sum_j |\kappa_j^{(t)}|\|\boldsymbol{a}_j\|}, & \text{if } \kappa_i^{(t)} \neq 0 \\ 0, & \text{otherwise} \end{cases}$
- *importance* - sample with a fixed non-uniform variant of *adaptive* obtained by bounding $\kappa_i^{(t)}$ with $2L_i$: $p_i := \frac{L_i\|\boldsymbol{a}_i\|}{\sum_j L_j\|\boldsymbol{a}_j\|}$.
- *ada-gap* - $p_i^{(t)} := \frac{G_i(\boldsymbol{\alpha}^{(t)})}{G(\boldsymbol{\alpha}^{(t)})}$.
- *gap-per-epoch* - Use *ada-gap* but with updates per-epoch. The gap-based distribution is only recomputed at the beginning of each epoch and stays fixed during each epoch.

## Experimental results

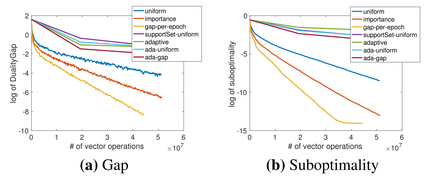



**(a)** Gap, fixed distr. **(b)** Subopt., fixed distr. **(c)** Gap, adaptive **(d)** Suboptimality, adaptive **(e)** Gap, fixed distr. **(f)** Subopt., fixed distr. **(g)** Gap, adaptive **(h)** Suboptimality, adaptive

**Figure 1:** Lasso (two top rows) and SVM (two bottom rows). Comparison of different fixed and adaptive variants of CD, reporting duality gap and suboptimality measures vs. epochs - mushrooms and ionosphere datasets.

| Dataset | $d$ | $n$ | nnz$/(nd)$ | $c_U = \frac{\mu(\|\boldsymbol{a}_i\|)}{\sigma(\|\boldsymbol{a}_i\|)}$ |
|---|---|---|---|---|
| mushrooms | 112 | 8124 | 18.8% | 1.34 |
| ionosphere | 351 | 33 | 88% | 3.07 |

**Table 1:** Datasets




**(a)** Gap **(b)** Suboptimality

**Figure 2:** Lasso on the *mushrooms* dataset. Performance in terms of duality gap and suboptimality, plotted against the total number of vector operations.

| Algorithm | Cost per Epoch |
|---|---|
| uniform | $\mathcal{O}(\text{nnz})$ |
| importance | $\mathcal{O}(\text{nnz} + n\log(n))$ |
| gap-per-epoch | $\mathcal{O}(\text{nnz} + n\log(n))$ |
| supportSet-uniform | $\mathcal{O}(n \cdot \text{nnz})$ |
| adaptive | $\mathcal{O}(n \cdot \text{nnz})$ |
| ada-uniform | $\mathcal{O}(n \cdot \text{nnz})$ |
| ada-gap | $\mathcal{O}(n \cdot \text{nnz})$ |

**Table 2:** A summary of computational costs

## Conclusion

- We investigated *adaptive* rules for adjusting the sampling probabilities in coordinate descent.
- Our theoretical results provide improved convergence rates for a more general class of algorithm schemes on one hand, and optimization problems on the other hand, where we are able to directly analyze CD on general convex objectives (as opposed to strongly convex regularizers in previous works).
- Our results are particularly useful for L1 problems and (original) hinge-loss objectives.
- We advocate the use of the computationally efficient gap-per-epoch sampling scheme in practice. While the scheme is close to the ones supported by our theory, an explicit primal-dual convergence analysis remains a future research question.

## References

[Csiba et al., 2015] 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 et al., 2016] 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 et al., 2016] 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.

## Acknowledgements

We thank Dominik Csiba and Peter Richtárik for helpful discussions. VC was supported in part by the European Commission under Grant ERC Future Proof, SNF 200021-146750, and SNF CRSII2-147633.