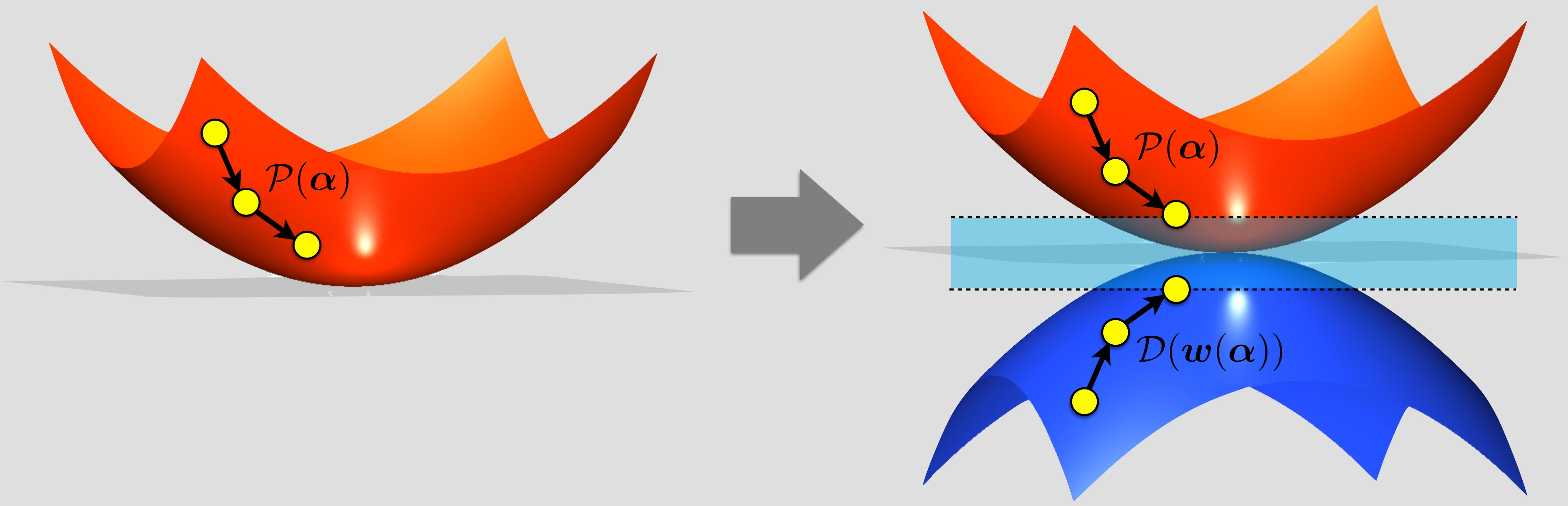
# Primal-Dual Rates and Certificates

Celestine Dünner [a,b] · Simone Forte [b] · Martin Takáč [c] · Martin Jaggi [b]

IBM [a] ETH [b] [c]

## overview

**Challenge:** most existing optimization algorithms lack accuracy certificates.




**Contributions**

- ✔ equip existing optimizers with accuracy **certificates**
- ✔ algorithm-independent new **primal-dual convergence rates** for a larger problem class

## setup

Problem formulation

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad f(A\boldsymbol{\alpha}) \;+\; g(\boldsymbol{\alpha})$$

gap

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad f^*(\boldsymbol{w}) \;+\; g^*(-A^\top \boldsymbol{w})$$

$f$ convex, smooth $\quad g$ convex

*correspondence mapping*
$\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{\alpha}) := \nabla f(A\boldsymbol{\alpha})$

Optimality conditions

$\mathbf{w} \in \partial f(A\boldsymbol{\alpha})$ $\quad -A^\top \mathbf{w} \in \partial g(\boldsymbol{\alpha})$ ,

$A\boldsymbol{\alpha} \in \partial f^*(\mathbf{w})$ $\quad \boldsymbol{\alpha} \in \partial g^*(-A^\top \mathbf{w})$

Convex conjugate: $\quad h^*(\boldsymbol{v}) := \sup_{\boldsymbol{u}\in\mathbb{R}^d} \; \boldsymbol{v}^T\boldsymbol{u} - h(\boldsymbol{u})$

## main results

**existing rates ⇒ primal-dual rates**

- algorithm agnostic

$g$ strongly convex

- linear rate ⇒ linear primal-dual rate

$f$ = L2, $g$ separable: see SDCA

$g$ bounded support

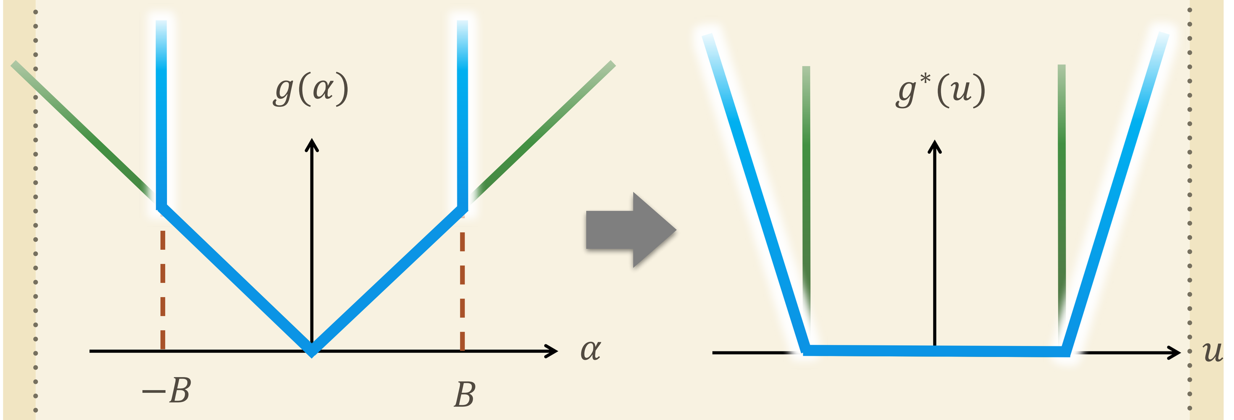
- linear rate ⇒ linear primal-dual rate
  new: SVM
- $1/T$ rate ⇒ $\sqrt{1/T}$ primal-dual rate

$g$ general convex?

- same! using trick, see next

*examples:* L1, elastic-n, group lasso, TV, fused L1, structured

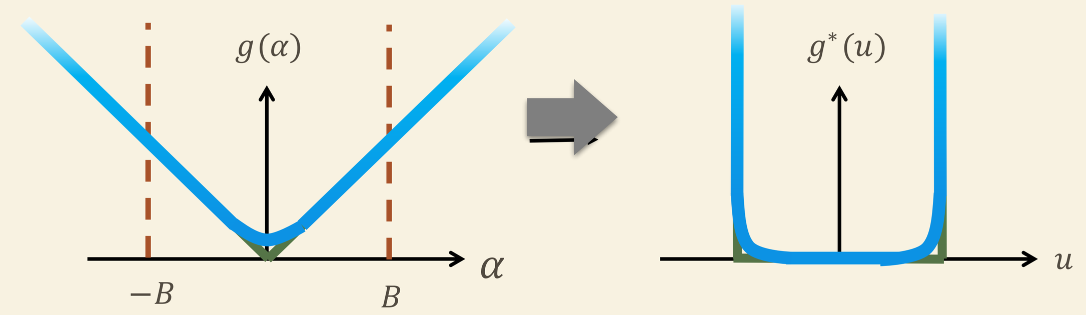
## lipschitzing trick




- makes $g^*$ globally Lipschitz
- gives duality gap defined on entire region of interest
  B easy to choose for norm-reg. problems
- problem and algorithms unaffected!

- *can re-use all existing algorithms!*

**Comparison to Nesterov Smoothing**




- makes $g^*$ strongly convex
- changes iterates

## proof details

**Lemma 1.** *Consider an optimization problem of the form (A). Let $f$ be $1/\beta$-smooth w.r.t. a norm $\|.\|_f$ and let $g$ be $\mu$-strongly convex with convexity parameter $\mu \geq 0$ w.r.t. a norm $\|.\|_g$. The general convex case $\mu = 0$ is explicitly allowed, but only if $g$ has bounded support.*

*Then, for any $\boldsymbol{\alpha} \in \mathrm{dom}(\mathcal{D})$ and any $s \in [0,1]$, it holds that*

$$\mathcal{D}(\boldsymbol{\alpha}) - \mathcal{D}(\boldsymbol{\alpha}^\star) \geq sG(\boldsymbol{\alpha}) + \tfrac{s^2}{2}\Big(\tfrac{\mu(1-s)}{s}\|\mathbf{u}-\boldsymbol{\alpha}\|_g^2 - \tfrac{1}{\beta}\|A(\mathbf{u}-\boldsymbol{\alpha})\|_f^2\Big) \tag{5}$$

*where $G(\boldsymbol{\alpha})$ is the gap function defined in (4) and*

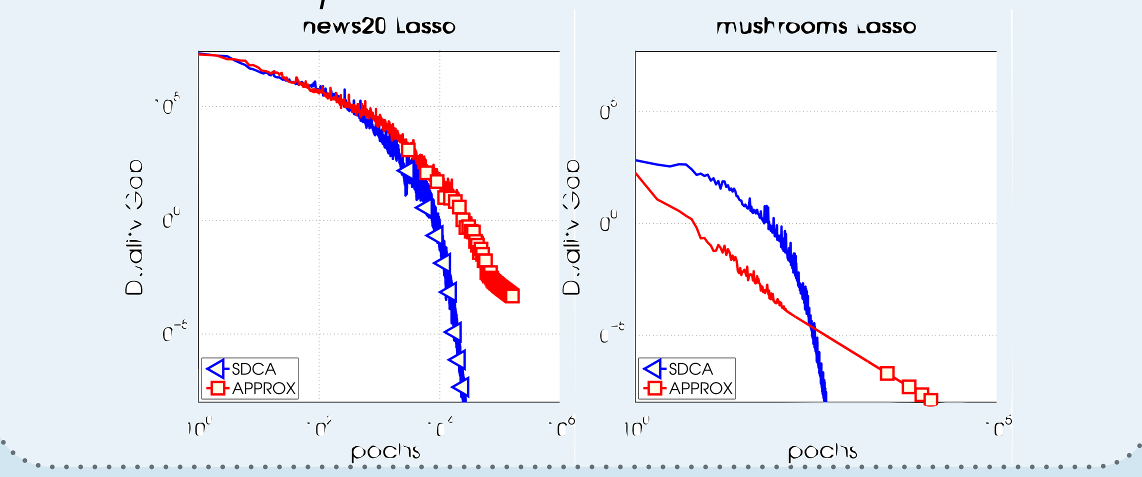
$$\mathbf{u} \in \partial g^*(-A^\top \mathbf{w}(\boldsymbol{\alpha})). \tag{6}$$

## coordinate descent

$f(A\boldsymbol{\alpha}) + \sum_i g_i(\alpha_i)$
$f^*(\mathbf{w}) + \sum_i g_i^*(-A_{:i}^\top \mathbf{w})$

- new primal-dual rates for CD on *L1, group-lasso, elastic-net*, etc
- certificates
- generalize SDCA. no square root

*illustrative experiments for L1*

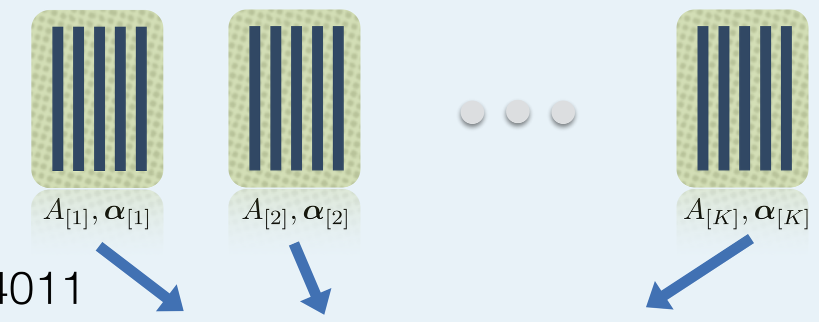



## distributed version

ML Systems Workshop on friday

**L1-Regularized Distributed Optimization: A Communication-Efficient Primal-Dual Framework**
Smith, V., Forte, S., Jordan, M. I., & Jaggi, M..

arXiv 1512.04011




## references

[1] Shalev-Shwartz and Zhang. *Stochastic dual coordinate ascent methods for regularized loss minimization.* JMLR, 14:567–599, 2013

[2] Necoara, I. (2015). Linear convergence of first order methods under weak nondegeneracy assumptions for convex programming. arXiv/1504.06298