Stochastic Dual Coordinate Ascent with Adaptive Probabilities

Dominik Csiba[1] Zheng Qu[1] Peter Richtárik[1]

[1]University of Edinburgh

Optimization and Big Data 2015
6. - 8. May, Edinburgh

Motivation

Empirical Risk Minimization

- Object-label pairs $(A_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ appear naturally in the world with unknown distribution $\mathcal{D}$.

Motivation

Empirical Risk Minimization

- Object-label pairs $(A_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ appear naturally in the world with unknown distribution $\mathcal{D}$.
- Find a vector $w \in \mathbb{R}^d$ such that for $(A_i, y_i) \sim \mathcal{D}$ we get

$$A_i^\top w \;\approx\; y_i.$$

Motivation

Empirical Risk Minimization

- Object-label pairs $(A_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ appear naturally in the world with unknown distribution $\mathcal{D}$.
- Find a vector $w \in \mathbb{R}^d$ such that for $(A_i, y_i) \sim \mathcal{D}$ we get

$$A_i^\top w \;\approx\; y_i.$$

- More precisely, we wish to find w solving

$$\min_w \mathbf{E}_{(A_i, y_i)\sim\mathcal{D}} \left[loss(A_i^T w, y_i) \right]$$

Motivation

Empirical Risk Minimization

- Object-label pairs $(A_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ appear naturally in the world with unknown distribution $\mathcal{D}$.
- Find a vector $w \in \mathbb{R}^d$ such that for $(A_i, y_i) \sim \mathcal{D}$ we get

$$A_i^\top w \ \approx \ y_i.$$

- More precisely, we wish to find w solving

$$\min_w \mathbf{E}_{(A_i, y_i) \sim \mathcal{D}} \left[loss(A_i^T w, y_i) \right]$$

1. Draw sample pairs $(A_i, y_i)_{i=1}^n$ from $\mathcal{D}$.

Motivation

Empirical Risk Minimization

- Object-label pairs $(A_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ appear naturally in the world with unknown distribution $\mathcal{D}$.
- Find a vector $w \in \mathbb{R}^d$ such that for $(A_i, y_i) \sim \mathcal{D}$ we get

$$A_i^\top w \;\approx\; y_i.$$

- More precisely, we wish to find w solving

$$\min_w \mathbf{E}_{(A_i,y_i)\sim\mathcal{D}} \left[loss(A_i^T w, y_i) \right]$$

1. Draw sample pairs $(A_i, y_i)_{i=1}^n$ from $\mathcal{D}$.
2. Take the empirical average

$$\min_w \frac{1}{n} \sum_{i=1}^{n} loss(A_i^\top w, y_i)$$

Motivation

Problem

Primal

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(A_i^\top w) + \lambda g(w) \right]$$

Primal

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(A_i^\top w) + \lambda g(w) \right]$$

Dual

$$\max_{\alpha \in \mathbb{R}^n} \left[D(\alpha) = -\lambda g^* \left(\frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i \right) - \frac{1}{n} \sum_{i=1}^{n} \phi_i^*(-\alpha_i) \right]$$

Main Contributions

- Two new algorithms

Main Contributions

- Two new algorithms
  - AdaSDCA Theoretical

Main Contributions

- Two new algorithms
  - AdaSDCA Theoretical
  - AdaSDCA+ Efficient variant of AdaSDCA

Main Contributions

- Two new algorithms
  - AdaSDCA Theoretical
  - AdaSDCA+ Efficient variant of AdaSDCA
- Properties

Main Contributions

- Two new algorithms
  - AdaSDCA Theoretical
  - AdaSDCA+ Efficient variant of AdaSDCA
- Properties
  - Coordinate descent on dual variables (SDCA-type algorithm)

Main Contributions

- Two new algorithms
  - AdaSDCA Theoretical
  - AdaSDCA+ Efficient variant of AdaSDCA
- Properties
  - Coordinate descent on dual variables (SDCA-type algorithm)
  - Adaptive probability distribution over dual coordinates

Main Contributions

- Two new algorithms
  - AdaSDCA Theoretical
  - AdaSDCA+ Efficient variant of AdaSDCA
- Properties
  - Coordinate descent on dual variables (SDCA-type algorithm)
  - Adaptive probability distribution over dual coordinates
  - First convergence guarantee for adaptive probability distribution

Main Contributions

- Two new algorithms
  - AdaSDCA Theoretical
  - AdaSDCA+ Efficient variant of AdaSDCA
- Properties
  - Coordinate descent on dual variables (SDCA-type algorithm)
  - Adaptive probability distribution over dual coordinates
  - First convergence guarantee for adaptive probability distribution
- Convergence Rate

Main Contributions

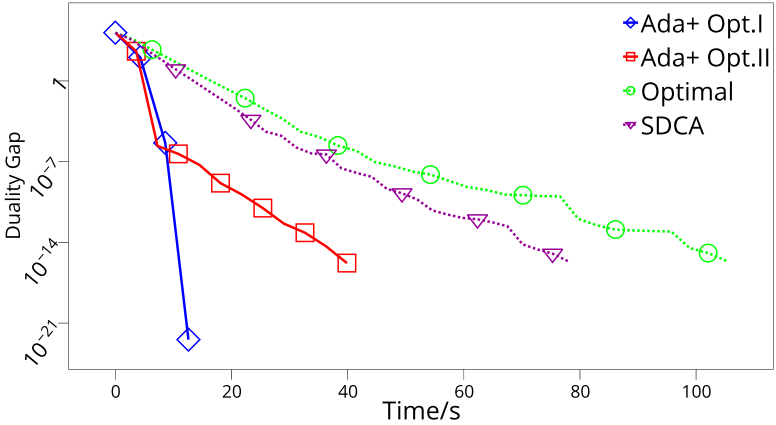
- Two new algorithms
  - AdaSDCA Theoretical
  - AdaSDCA+ Efficient variant of AdaSDCA
- Properties
  - Coordinate descent on dual variables (SDCA-type algorithm)
  - Adaptive probability distribution over dual coordinates
  - First convergence guarantee for adaptive probability distribution
- Convergence Rate
  - AdaSDCA enjoys better rate than the best known rate for SDCA with importance sampling

Importance Sampling

$$T \geq \left(n + \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\lambda\gamma} \right) \log\left(\frac{C}{\epsilon}\right) \Rightarrow \mathbb{E}[P(w^T) - D(\alpha^T)] \leq \epsilon$$

Experiments

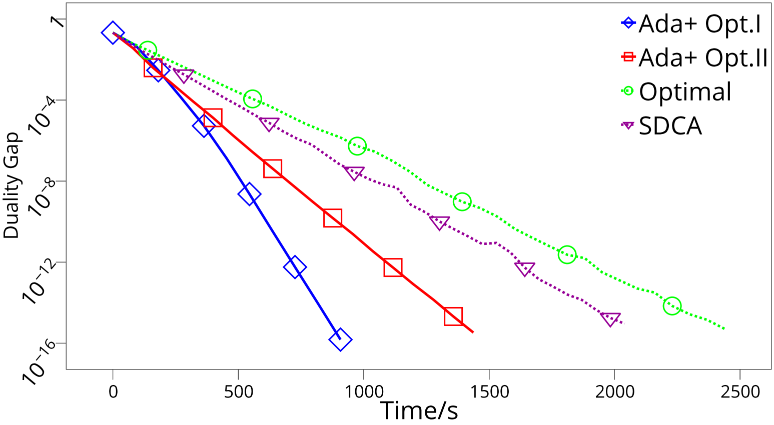
cov1 dataset, $d = 54, n = 581,012$



Smooth Hinge loss with L_2 regularizer

Experiments

synthetic dataset, $d = 100$, $n = 10,000,000$, sparsity$= 0.1$



Smooth Hinge loss with L_2 regularizer

Thank you for your attention!