Stochastic Dual Coordinate Ascent with Adaptive Probabilities

Dominik Csiba

University of Edinburgh

Trojkráľová konferencia
4 January, Bratislava

Co-authors

Zheng Qu

Peter Richtárik

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

Motivation
Empirical Risk Minimization

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

$$x_i^\top w \approx y_i$$

Motivation
Empirical Risk Minimization

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

$$x_i^\top w \approx y_i$$

- More precisely, we wish to find w solving

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

Motivation
Empirical Risk Minimization

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

$$x_i^\top w \approx y_i$$

- More precisely, we wish to find w solving

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

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

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

$$x_i^\top w \approx y_i$$

- More precisely, we wish to find w solving

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

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

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

Empirical Risk Minimization

Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

Empirical Risk Minimization

Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

- supervised learning

Empirical Risk Minimization

Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

- supervised learning
- train a linear predictor $w \in \mathbb{R}^d$

Empirical Risk Minimization

Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

- supervised learning
- train a linear predictor $w \in \mathbb{R}^d$
- n training samples $x_1, \ldots, x_n \in \mathbb{R}^d$

Empirical Risk Minimization

Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

- supervised learning
- train a linear predictor $w \in \mathbb{R}^d$
- n training samples $x_1, \ldots, x_n \in \mathbb{R}^d$
- convex and $1/\gamma$-smooth loss function $\phi_i : \mathbb{R} \to \mathbb{R}$

Empirical Risk Minimization

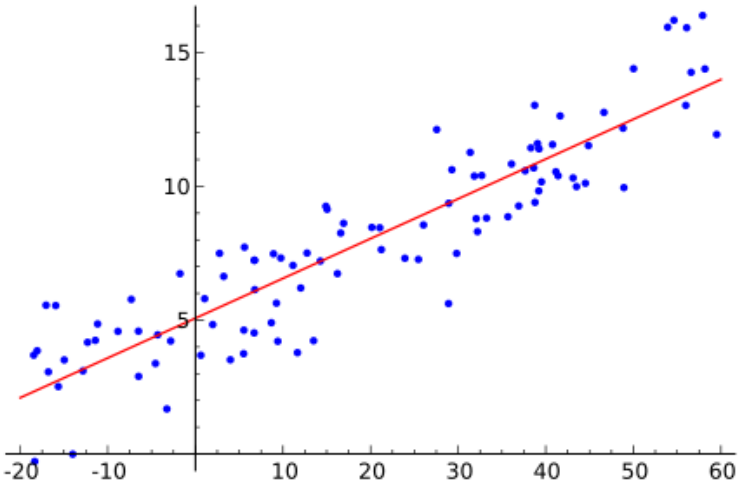
Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

- supervised learning
- train a linear predictor $w \in \mathbb{R}^d$
- n training samples $x_1, \ldots, x_n \in \mathbb{R}^d$
- convex and $1/\gamma$-smooth loss function $\phi_i : \mathbb{R} \to \mathbb{R}$
  - ex.: Squared loss $(\phi_i(a) = \frac{1}{2\gamma}(a - y_i)^2)$,
    Logistic loss $(\phi_i(a) = \frac{4}{\gamma} \log(1 + e^{-y_i a}))$, ...

Empirical Risk Minimization

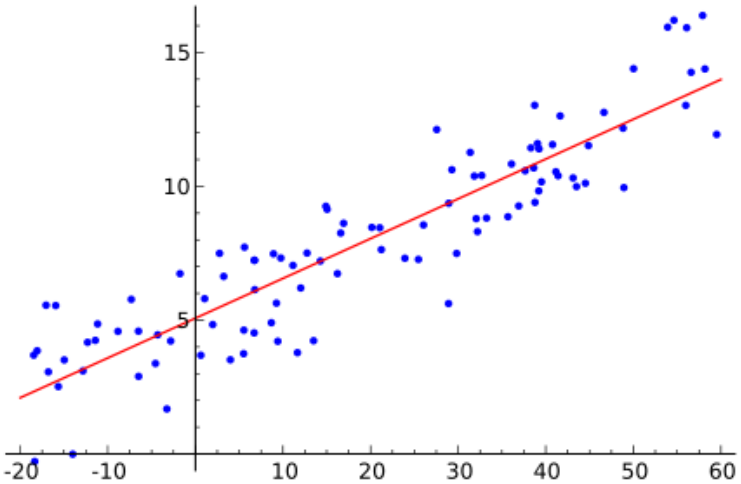
Linear Regression



$$\phi_i(u) = \frac{1}{2}(u - y_i)^2$$

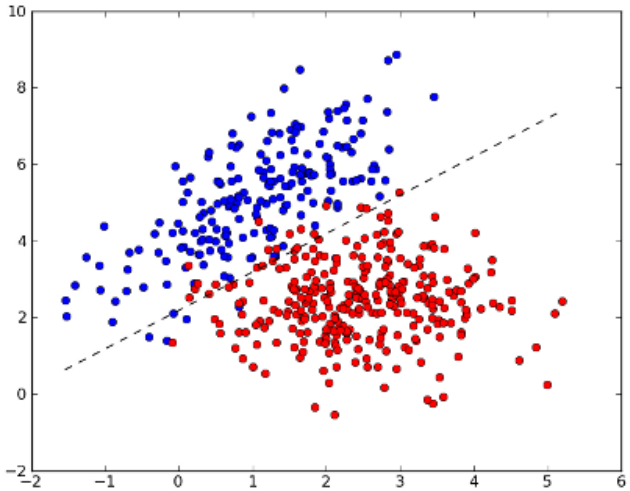
Empirical Risk Minimization

Linear Regression



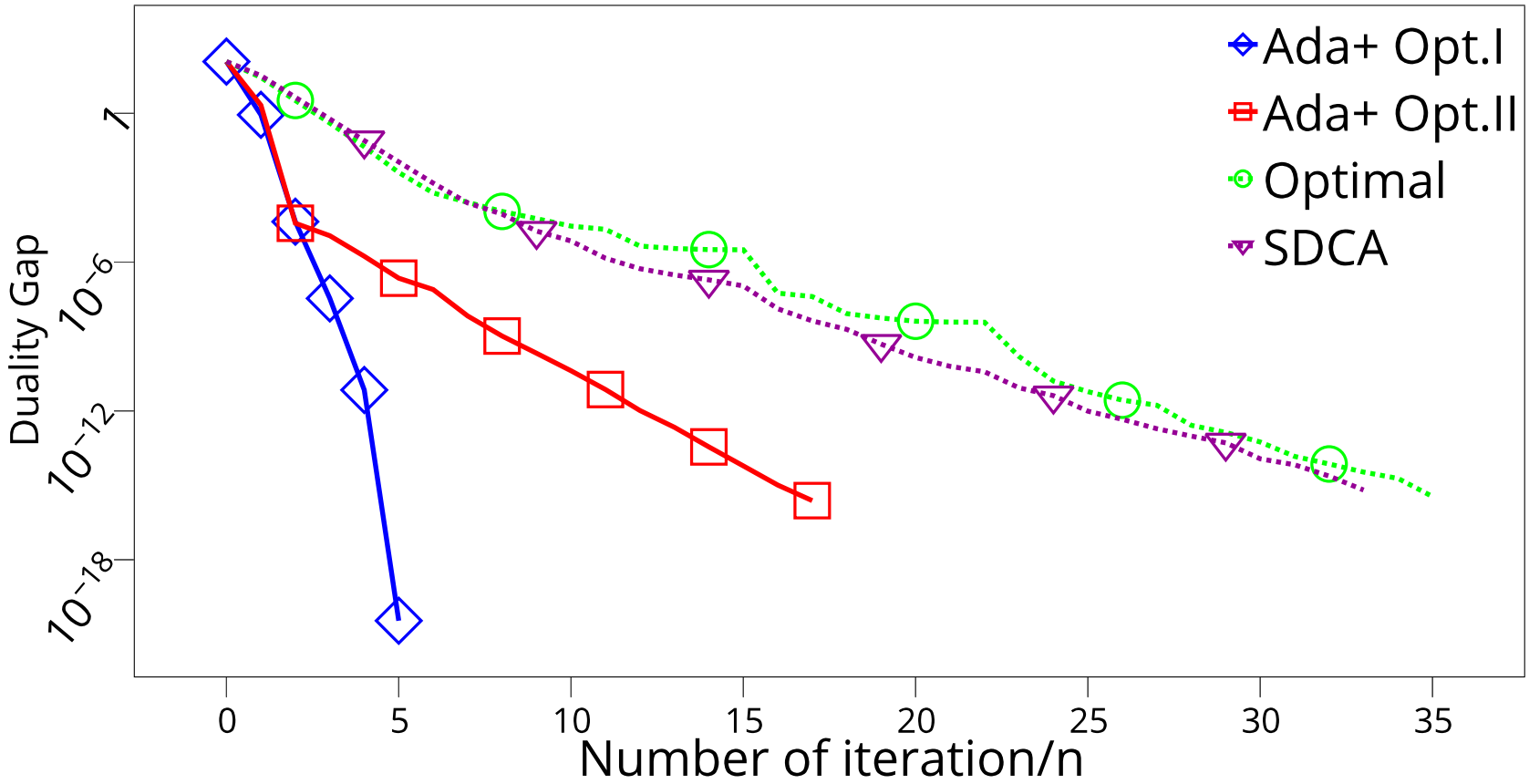
$$\phi_i(u) = \frac{1}{2}(u - y_i)^2$$

Logistic Regression



$$\phi_i(u) = \log(1 + \exp(-y_i u))$$

A picture is worth a thousand words



Primal Dual Formulation

- Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

Primal Dual Formulation

- Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

- Dual problem of the regularized ERM:

$$\max_{\alpha \in \mathbb{R}^n} \; D(\alpha) \stackrel{\text{def}}{=} \underbrace{-\frac{1}{2\lambda n^2} \left\| \sum_{i=1}^{n} x_i \alpha_i \right\|_2^2}_{\text{quadratic}} \underbrace{- \frac{1}{n} \sum_{i=1}^{n} \phi_i^\star(-\alpha_i)}_{\text{strongly convex and separable}}$$

Primal Dual Formulation

- Regularized ERM:

$$\min_{w\in\mathbb{R}^d}\left[P(w)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}\phi_i(x_i^\top w)+\frac{\lambda}{2}\|w\|_2^2\right]$$

- Dual problem of the regularized ERM:

$$\max_{\alpha\in\mathbb{R}^n}\ D(\alpha)\overset{\text{def}}{=}\underbrace{-\frac{1}{2\lambda n^2}\left\|\sum_{i=1}^{n}x_i\alpha_i\right\|_2^2}_{\text{quadratic}}\underbrace{-\frac{1}{n}\sum_{i=1}^{n}\phi_i^\star(-\alpha_i)}_{\text{strongly convex and separable}}$$

- Optimality conditions:

$$\textbf{OPT1}:\ w^*=\frac{1}{\lambda n}\sum_{i=1}^{n}x_i\alpha_i^*$$

$$\textbf{OPT2}:\ \alpha_i^*=-\nabla\phi_i(x_i^\top w^*),\ \ \forall i=1,\dots,n.$$

Optimality Conditions

- Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

Optimality Conditions

- Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

$$\vec{0} = \nabla P(w^*) = \frac{1}{n} \sum_{i=1}^{n} x_i \nabla \phi_i(x_i^\top w^*) + \lambda w^*$$

$$w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^*,$$

where $\alpha_i^* \stackrel{\text{def}}{=} -\nabla \phi_i(x_i^\top w^*)$

Optimality Conditions

- Regularized ERM:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi_i(x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

$$\vec{0} = \nabla P(w^*) = \frac{1}{n} \sum_{i=1}^{n} x_i \nabla \phi_i(x_i^\top w^*) + \lambda w^*$$

$$\textbf{OPT1}: \; w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^*$$

$$\textbf{OPT2}: \; \alpha_i^* = -\nabla \phi_i(x_i^\top w^*) \quad \forall i = 1, \ldots, n$$

Stochastic Dual Coordinate Ascent

$$\textbf{OPT1}: \; w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^*$$

$$\textbf{OPT2}: \; \alpha_i^* = -\nabla \phi_i(x_i^\top w^*), \;\; \forall i = 1, \ldots, n.$$

Algorithm SDCA

1: **for** $t = 1, \ldots$ **do**
2: Update w^{t+1} according to **OPT1**
3: Randomly sample $i \in 1, \ldots, n$
4: Update α_i^{t+1} according to **OPT2**
5: **until happy**
6: **end for**

Stochastic Dual Coordinate Ascent

$$\textbf{OPT1}: \; w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^*$$

$$\textbf{OPT2}: \; \alpha_i^* = -\nabla \phi_i(x_i^\top w^*), \;\; \forall i = 1, \ldots, n.$$

Algorithm SDCA

1: **for** $t = 1, \ldots$ **do**
2: $\quad w^{t+1} = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^t$
3: $\quad$ Randomly sample $i \in 1, \ldots, n$
4: $\quad \alpha_i^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_i^\star(-\beta) - (x_i^\top w^t)\beta - \frac{\|x_i\|^2}{2\lambda n} |\beta - \alpha_i^t|^2 \right\}$
5: $\quad$ **until happy**
6: **end for**

Stochastic Dual Coordinate Ascent

$$\textbf{OPT1}: \; w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^*$$

$$\textbf{OPT2}: \; \alpha_i^* = -\nabla \phi_i(x_i^\top w^*), \;\; \forall i = 1, \ldots, n.$$

Algorithm SDCA

1: **for** $t = 1, \ldots$ **do**
2: $\quad w^{t+1} = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^t$
3: $\quad$ Randomly sample $i \in 1, \ldots, n$
4: $\quad \alpha_i^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_i^\star(-\beta) - (x_i^\top w^t)\beta - \frac{\|x_i\|^2}{2\lambda n} |\beta - \alpha_i^t|^2 \right\}$
5: $\quad$ **until happy**
6: **end for**

Uniform and Importance Sampling

Uniform sampling (SDCA: [Shalev-Shwartz & Zhang 13'],...)

$$p_i = \textbf{Prob}(i \text{ is sampled}) \sim \frac{1}{n},$$

Iteration complexity:

$$\tilde{O}\left(n + \frac{\max_i \|x_i\|^2}{\lambda\gamma}\right)$$

Uniform and Importance Sampling

Uniform sampling (SDCA: [Shalev-Shwartz & Zhang 13'],...)

$$p_i = \textbf{Prob}(i \text{ is sampled}) \sim \frac{1}{n},$$

Iteration complexity:

$$\tilde{O}\left(n + \frac{\max_i \|x_i\|^2}{\lambda\gamma}\right)$$

Importance sampling (Iprox-SDCA: [Zhao & Zhang 15'],...)

$$p_i = \textbf{Prob}(i \text{ is sampled}) \sim \|x_i\|^2 + \lambda\gamma n,$$

Iteration complexity:

$$\tilde{O}\left(n + \frac{\frac{1}{n}\sum_{i=1}^n \|x_i\|^2}{\lambda\gamma}\right)$$

Stochastic Dual Coordinate Ascent

Algorithm SDCA

1: **for** $t = 1, \ldots$ **do**
2: $w^{t+1} = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^t$
3: Randomly sample $i \in 1, \ldots, n$ according to a fixed distribution p
4: $\alpha_i^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_i^\star(-\beta) - (x_i^\top w^t)\beta - \frac{\|x_i\|^2}{2\lambda n} |\beta - \alpha_i^t|^2 \right\}$
5: **until happy**
6: **end for**

Adaptive Stochastic Dual Coordinate Ascent

Algorithm AdaSDCA

1: **for** $t = 1, \ldots$ **do**
2: $\quad w^{t+1} = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^t$
3: $\quad$ Randomly sample $i \in 1, \ldots, n$ according to p^t
4: $\quad \alpha_i^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_i^\star(-\beta) - (x_i^\top w^t)\beta - \frac{\|x_i\|^2}{2\lambda n} |\beta - \alpha_i^t|^2 \right\}$
5: $\quad$ Change/adapt the probability distribution p^t to get p^{t+1}
6: $\quad$ **until happy**
7: **end for**

Adaptive Stochastic Dual Coordinate Ascent

Algorithm AdaSDCA

1: **for** $t = 1, \ldots$ **do**
2: $\quad w^{t+1} = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^t$
3: $\quad$ Randomly sample $i \in 1, \ldots, n$ according to p^t
4: $\quad \alpha_i^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_i^\star(-\beta) - (x_i^\top w^t)\beta - \frac{\|x_i\|^2}{2\lambda n} |\beta - \alpha_i^t|^2 \right\}$
5: $\quad$ Change/adapt the probability distribution p^t to get p^{t+1}
6: $\quad$ **until happy**
7: **end for**

How to change the distribution?

Dual Residue

- **Measure of progress:** based on

$$\textbf{OPT2}: \ \alpha_i^* = -\nabla\phi_i(x_i^\top w^*), \ \ \forall i \in [n].$$

we introduce the dual residue κ_i^t

$$\kappa_i^t \overset{\text{def}}{=} \alpha_i^t + \nabla\phi_i(x_i^\top w^t), \quad i = 1, \ldots, n$$

Dual Residue

- **Measure of progress:** based on

$$\textbf{OPT2}: \ \alpha_i^* = -\nabla\phi_i(x_i^\top w^*), \ \ \forall i \in [n].$$

we introduce the dual residue κ_i^t

$$\kappa_i^t \overset{\text{def}}{=} \alpha_i^t + \nabla\phi_i(x_i^\top w^t), \quad i = 1, \ldots, n$$

- **Note:** at optimality, $\kappa_i^t = 0, \quad \forall i = 1, \ldots, n$

Optimal Adaptive Sampling

- Let

$$p_i^t \stackrel{\text{def}}{=} \textbf{Prob}(i \text{ is sampled at iteration } t)$$

Optimal Adaptive Sampling

- Let

$$p_i^t \stackrel{\text{def}}{=} \textbf{Prob}(i \text{ is sampled at iteration } t)$$

- According to the theory, the best adaptive sampling is

$$p_i^t \sim |\kappa_i^t| \sqrt{\|x_i\|_2^2 + n\lambda\gamma}$$

Optimal Adaptive Sampling

- Let

$$p_i^t \stackrel{\text{def}}{=} \textbf{Prob}(i \text{ is sampled at iteration } t)$$

- According to the theory, the best adaptive sampling is

$$p_i^t \sim |\kappa_i^t| \sqrt{\|x_i\|_2^2 + n\lambda\gamma}$$

- **Issue:** calculating κ_i^t and sampling according to it at every iteration is very costly

Adaptive Stochastic Dual Coordinate Ascent

Optimal algorithm:

Algorithm AdaSDCA

1: **for** $t = 1, \ldots$ **do**
2: $w^{t+1} = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^t$
3: Randomly sample $i \in 1, \ldots, n$ according to p^t
4: $\alpha_i^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_i^\star(-\beta) - (x_i^\top w^t)\beta - \frac{\|x_i\|^2}{2\lambda n} |\beta - \alpha_i^t|^2 \right\}$
5: Set $p^{t+1} \sim |\kappa_i^t| \sqrt{\|x_i\|_2^2 + n\lambda\gamma}$
6: **until happy**
7: **end for**

Adaptive Stochastic Dual Coordinate Ascent

Optimal algorithm:

Algorithm AdaSDCA

1: **for** $t = 1, \ldots$ **do**
2: $w^{t+1} = \frac{1}{\lambda n} \sum_{i=1}^{n} x_i \alpha_i^t$
3: Randomly sample $i \in 1, \ldots, n$ according to p^t
4: $\alpha_i^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_i^\star(-\beta) - (x_i^\top w^t)\beta - \frac{\|x_i\|^2}{2\lambda n} |\beta - \alpha_i^t|^2 \right\}$
5: Set $p^{t+1} \sim |\kappa_i^t| \sqrt{\|x_i\|_2^2 + n\lambda\gamma}$
6: **until happy**
7: **end for**

... but infeasible to implement.

Efficient heuristic implementation AdaSDCA+:

Algorithm AdaSDCA+

for $t = 1, \ldots$ **do**
  if $\mathrm{mod}(t, n) == 1$ **then**
    Set $p^t \sim |\kappa_i^t| \sqrt{\|x_i\|_2^2 + n\lambda\gamma}$
  end if
  $w^{t+1} = \frac{1}{\lambda n} \sum_{i=1}^n x_i \alpha_i^t$
  Randomly sample $i \in 1, \ldots, n$ according to p^t
  $\alpha_i^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_i^\star(-\beta) - (x_i^\top w^t)\beta - \frac{\|x_i\|^2}{2\lambda n} |\beta - \alpha_i^t|^2 \right\}$
  Set $p_i^{t+1} \sim p_i^{t+1}/m$
  until happy
end for

Computational Cost per Epoch

Algorithm	cost of an epoch
SDCA	$O(\text{nnz})$
IProx-SDCA	$O(\text{nnz} + n\log(n))$
AdaSDCA	$O(n \cdot \text{nnz})$
AdaSDCA+	$O(\text{nnz} + n\log(n))$

Table 1 : One epoch computational cost of different algorithms

Numerical Experiments

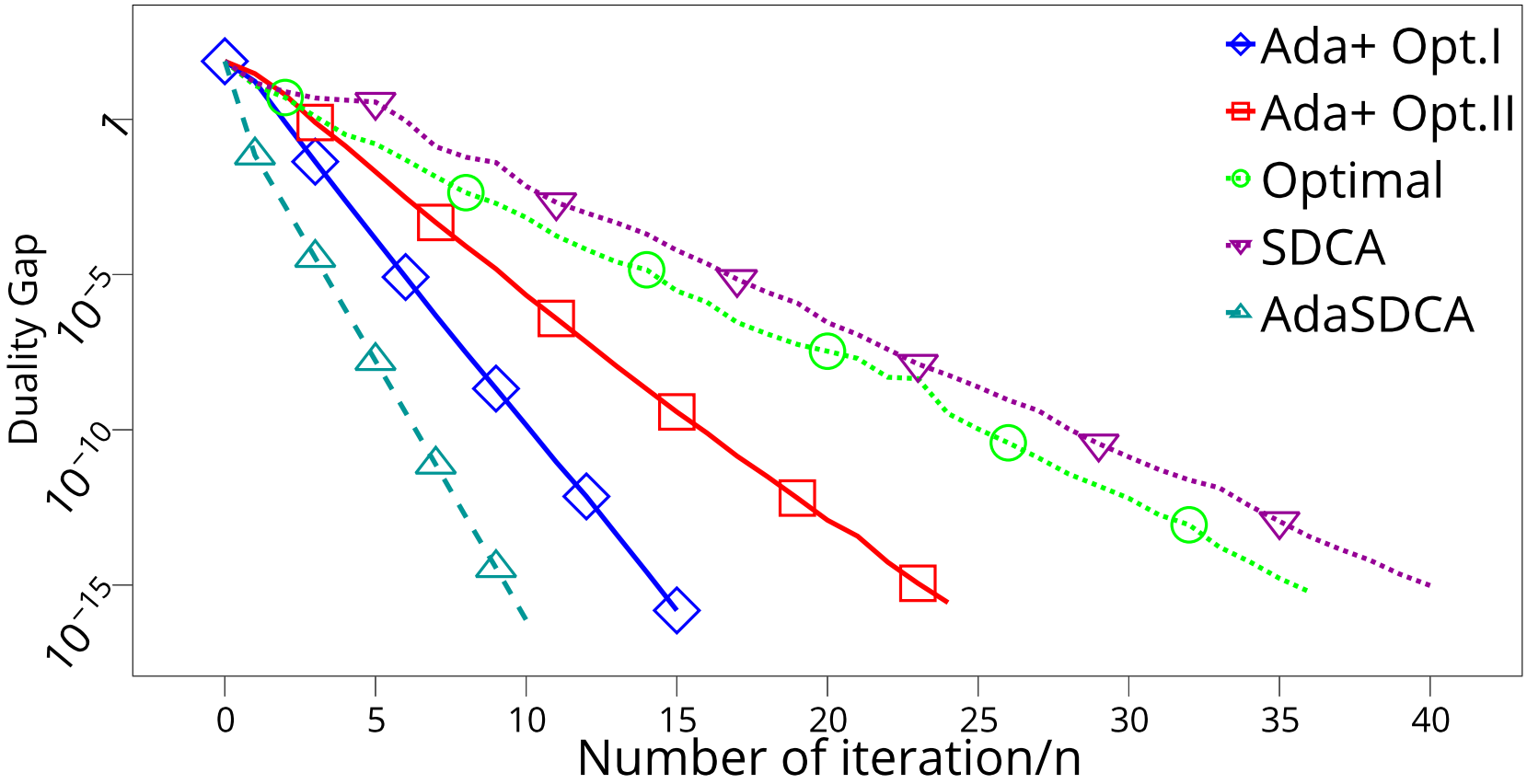


Figure 1 : **w8a** dataset $d = 300$, $n = 49749$.
Quadratic loss, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

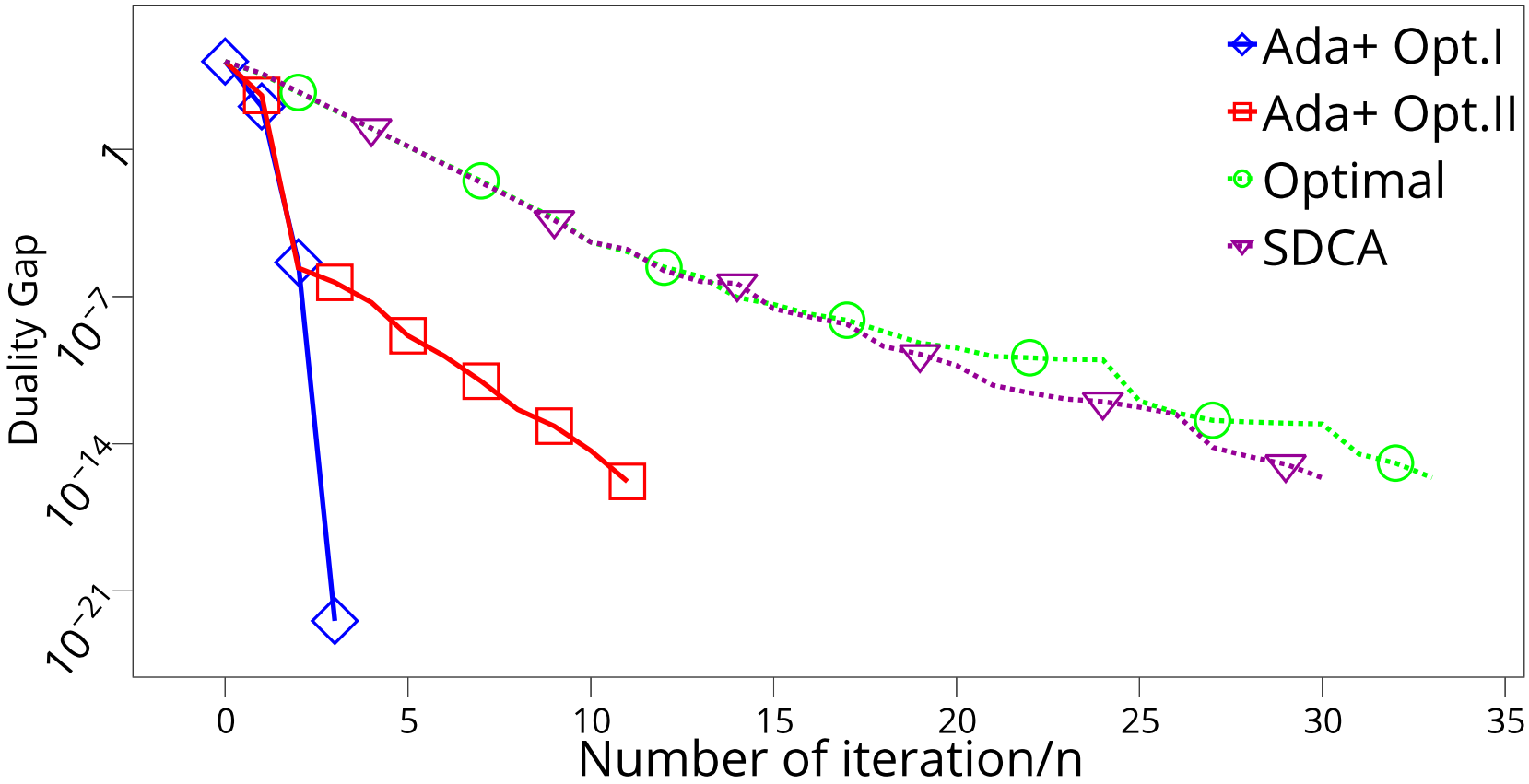


Figure 2 : **cov1** dataset: $d = 54, n = 581,012$.
Smooth Hinge loss, $\lambda = 1/n, \gamma = 1$.

Summary

- First analysis of an adaptive probability distribution

Summary

- First analysis of an adaptive probability distribution
- Theoretical method **AdaSDCA** beats the current state-of-the-art

Summary

- First analysis of an adaptive probability distribution
- Theoretical method **AdaSDCA** beats the current state-of-the-art
- Efficient heuristic implementation **AdaSDCA+**

Summary

- First analysis of an adaptive probability distribution
- Theoretical method **AdaSDCA** beats the current state-of-the-art
- Efficient heuristic implementation **AdaSDCA+**
- **Follow-up:** parallel version of **AdaSDCA** and **AdaSDCA+**

Thank you for your attention!