Stochastic Dual Coordinate Ascent with Adaptive Probabilities

Zheng Qu

The University of Hong Kong

The China-R Conference
27-29 May 2016, Beijing

Dominik Csiba, Zheng Qu and Peter Richtárik. *Stochastic Dual Coordinate Ascent with Adaptive Probabilities, ICML, 2015.*

Empirical Risk Minimization

ERM:

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

- supervised learning;
- train a linear predictor $w \in \mathbb{R}^d$;
- n training samples $A_1, \ldots, A_n \in \mathbb{R}^d$;

Empirical Risk Minimization

ERM:

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

- supervised learning;
- train a linear predictor $w\in\mathbb{R}^d$;
- n training samples $A_1,\ldots,A_n\in\mathbb{R}^d$;
- convex and $1/\gamma$-smooth loss function $\phi_i:\mathbb{R}\to\mathbb{R}$;
  - ex.: Squared loss $(\phi_i(a)=\frac{1}{2\gamma}(a-b_i)^2)$, Logistic loss $(\phi_i(a)=\frac{4}{\gamma}\log(1+e^a))$, ...

Empirical Risk Minimization

ERM:

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

- supervised learning;
- train a linear predictor $w\in\mathbb{R}^d$;
- n training samples $A_1,\ldots,A_n\in\mathbb{R}^d$;
- convex and $1/\gamma$-smooth loss function $\phi_i:\mathbb{R}\to\mathbb{R}$;
  - ex.: Squared loss $(\phi_i(a)=\frac{1}{2\gamma}(a-b_i)^2)$, Logistic loss $(\phi_i(a)=\frac{4}{\gamma}\log(1+e^a))$, ...
- 1-strongly convex regularizer $g:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$;
  - ex.: L_2 regularization $(g(w)=\frac{1}{2}\|w\|_2^2)$, ...

Primal Dual Formulation

- ERM:

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

- Dual problem of ERM:

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

Primal Dual Formulation

- ERM:

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

- Dual problem of ERM:

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

- Optimality conditions:

$$\textbf{OPT1}:\ w^*=\nabla g^\star\left(\frac{1}{\lambda n}A\alpha^*\right)$$

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

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \ldots, n\}$;**
3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^\star(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Stochastic Dual Coordinate Ascent

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \ldots, n\}$;**
3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^\star(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^\star(\frac{1}{\lambda n} A\alpha^t)$

Stochastic Dual Coordinate Ascent

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \ldots, n\}$ according to a fixed distribution p;**
3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^\star(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^\star(\frac{1}{\lambda n} A\alpha^t)$

Uniform and Importance Sampling

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

$$p_i = \mathbf{Prob}(i_t = i) \sim \frac{1}{n},$$

Iteration complexity:

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

Uniform and Importance Sampling

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

$$p_i = \mathbf{Prob}(i_t = i) \sim \frac{1}{n},$$

Iteration complexity:

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

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

$$p_i = \mathbf{Prob}(i_t = i) \sim \|A_i\|^2 + \lambda\gamma n,$$

Iteration complexity:

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

Adaptive Sampling

- Each dual variable has a natural measure of progress:

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

called dual residue.

Adaptive Sampling

- Each dual variable has a natural measure of progress:

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

  called dual residue.
- Optimality conditions:

$$\textbf{OPT1}: \ w^* = \nabla g^* \left(\frac{1}{\lambda n} A\alpha^* \right)$$

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

Adaptive Sampling

- Each dual variable has a natural measure of progress:

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

  called dual residue.
- Optimality conditions:

$$\textbf{OPT1}: \; w^* = \nabla g^* \left(\frac{1}{\lambda n} A\alpha^* \right)$$

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

- A sampling distribution p is coherent with κ^t if for all $i \in [n]$:

$$\kappa_i^t \neq 0 \quad \Rightarrow p_i > 0.$$

Stochastic Dual Coordinate Ascent

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \ldots, n\}$ according to a fixed distribution p;**
3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^\star(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^\star(\frac{1}{\lambda n} A\alpha^t)$

Adaptive Stochastic Dual Coordinate Ascent

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \ldots, n\}$ according to a distribution p^t coherent with dual residue κ^t;**
3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^\star(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^\star(\frac{1}{\lambda n} A\alpha^t)$

Convergence Theorem

Theorem (AdaSDCA)

Consider AdaSDCA. If at each iteration $t \geq 0$,

$$\theta(\kappa^t, p^t) \overset{def}{=} \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i^t)^{-1} (\|A_i\|^2 + n\lambda\gamma)|\kappa_i^t|^2} \leq \min_{i:\kappa_i^t \neq 0} p_i^t,$$

then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^{t} (1 - \tilde{\theta}_k) \left(D(\alpha^*) - D(\alpha^0) \right),$$

for all $t \geq 0$ where

$$\tilde{\theta}_t \overset{def}{=} \frac{\mathbb{E}[\theta(\kappa^t, p^t)(P(w^t) - D(\alpha^t))]}{\mathbb{E}[P(w^t) - D(\alpha^t)]}.$$

Optimal Adaptive Sampling Probability

$$
\begin{aligned}
p^*(\kappa^t) = \quad & \arg\max \quad \theta(\kappa^t, p) \\
& \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \textstyle\sum_i p_i = 1 \\
& \qquad\quad\ \ p \text{ is coherent with } \kappa^t \\
& \qquad\quad\ \ \theta(\kappa^t, p) \leq \min_{i:\kappa^t_i \neq 0} p_i
\end{aligned}
$$

Optimal Adaptive Sampling Probability

$$
\begin{aligned}
p^*(\kappa^t) = \quad & \arg\max \quad \theta(\kappa^t, p) \\
& \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \textstyle\sum_i p_i = 1 \\
& \qquad\quad\ p \text{ is coherent with } \kappa^t \\
& \qquad\quad\ \theta(\kappa^t, p) \leq \min_{i:\kappa^t_i \neq 0} p_i
\end{aligned}
$$

Relaxation:

$$
\begin{aligned}
\tilde{p}^*(\kappa^t) = \quad & \arg\max \quad \theta(\kappa^t, p) \\
& \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \textstyle\sum_{i=1}^n p_i = 1
\end{aligned}
$$

Optimal Adaptive Sampling Probability

$$
\begin{aligned}
p^*(\kappa^t) = \quad & \arg\max \quad \theta(\kappa^t, p) \\
& \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \textstyle\sum_i p_i = 1 \\
& \qquad\quad\ p \text{ is coherent with } \kappa^t \\
& \qquad\quad\ \theta(\kappa^t, p) \leq \min_{i:\kappa_i^t \neq 0} p_i
\end{aligned}
$$

Relaxation:

$$
\begin{aligned}
\tilde{p}^*(\kappa^t) = \quad & \arg\max \quad \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i)^{-1} |\kappa_i^t|^2 (\|A_i\|^2 + n\lambda\gamma)} \\
& \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \textstyle\sum_{i=1}^n p_i = 1
\end{aligned}
$$

Optimal Adaptive Sampling Probability

$$
\begin{aligned}
p^*(\kappa^t) = \quad & \arg\max \quad \theta(\kappa^t, p) \\
& \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \textstyle\sum_i p_i = 1 \\
& \qquad\quad\ p \text{ is coherent with } \kappa^t \\
& \qquad\quad\ \theta(\kappa^t, p) \leq \min_{i:\kappa^t_i \neq 0} p_i
\end{aligned}
$$

Relaxation:

$$
\begin{aligned}
\tilde{p}^*(\kappa^t) = \quad & \arg\max \quad \frac{n\lambda\gamma \sum_i |\kappa^t_i|^2}{\sum_{i:\kappa^t_i \neq 0} (p_i)^{-1} |\kappa^t_i|^2 (\|A_i\|^2 + n\lambda\gamma)} \\
& \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \textstyle\sum_{i=1}^n p_i = 1
\end{aligned}
$$

$$
(\tilde{p}^*(\kappa^t))_i \sim |\kappa^t_i| \sqrt{\|A_i\|^2 + n\lambda\gamma}, \quad \forall i \in [n].
$$

Exact Relaxation for Squared Loss

Theorem (AdaSDCA)

Consider AdaSDCA. If at each iteration $t \geq 0$,

$$\theta(\kappa^t, p^t) \overset{def}{=} \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i^t)^{-1} |\kappa_i^t|^2 (\|A_i\|^2 + n\lambda\gamma)} \leq \min_{i:\kappa_i^t \neq 0} p_i^t,$$

then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^{t} (1 - \tilde{\theta}_k) \left(D(\alpha^*) - D(\alpha^0) \right),$$

for all $t \geq 0$ where

$$\tilde{\theta}_t \overset{def}{=} \frac{\mathbb{E}[\theta(\kappa^t, p^t)(P(w^t) - D(\alpha^t))]}{\mathbb{E}[P(w^t) - D(\alpha^t)]}.$$

Exact Relaxation for Squared Loss

Theorem (AdaSDCA for squared loss)

Consider AdaSDCA. If all the loss functions $\{\phi_i\}$ are squared loss functions, then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^{t} (1 - \tilde{\theta}_k) \left(D(\alpha^*) - D(\alpha^0) \right),$$

for all $t \geq 0$ where

$$\tilde{\theta}_t \overset{def}{=} \frac{\mathbb{E}[\theta(\kappa^t, p^t)(P(w^t) - D(\alpha^t))]}{\mathbb{E}[P(w^t) - D(\alpha^t)]}.$$

Exact Relaxation for Squared Loss

Theorem (AdaSDCA for squared loss)

Consider AdaSDCA. If all the loss functions $\{\phi_i\}$ are squared loss functions, then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^{t} (1 - \tilde{\theta}_k) \left(D(\alpha^*) - D(\alpha^0) \right),$$

for all $t \geq 0$ where

$$\tilde{\theta}_t \overset{def}{=} \frac{\mathbb{E}[\theta(\kappa^t, p^t)(P(w^t) - D(\alpha^t))]}{\mathbb{E}[P(w^t) - D(\alpha^t)]}.$$

Optimal adaptive sampling probability is given by:

$$(\tilde{p}^*(\kappa^t))_i \sim |\kappa_i^t| \sqrt{\|A_i\|^2 + n\lambda\gamma}, \quad \forall i \in [n].$$

Heuristic and Efficient Variant of AdaSDCA

AdaSDCA+:

Dual solution

For $t \geq 1$:

1. If $mod(t, n) = 0$, then

   Option I: Adaptive Sampling Probability

   Compute dual residue κ^t: $\kappa_i^t = \alpha_i^t + \nabla\phi_i(A_i^\top w^t)$

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

   Option II: Importance Sampling Probability

   Set $p_i^t \sim \|A_i\|^2 + n\lambda\gamma$

2. Randomly pick $i_t \in \{1, \ldots, n\}$ according to p^t
3. Update $\alpha_{i_t}^t$
4. Update Probability: $p^{t+1} \sim (p_1^t, \ldots, p_{i_t}^t/m, \ldots p_n^t)$

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

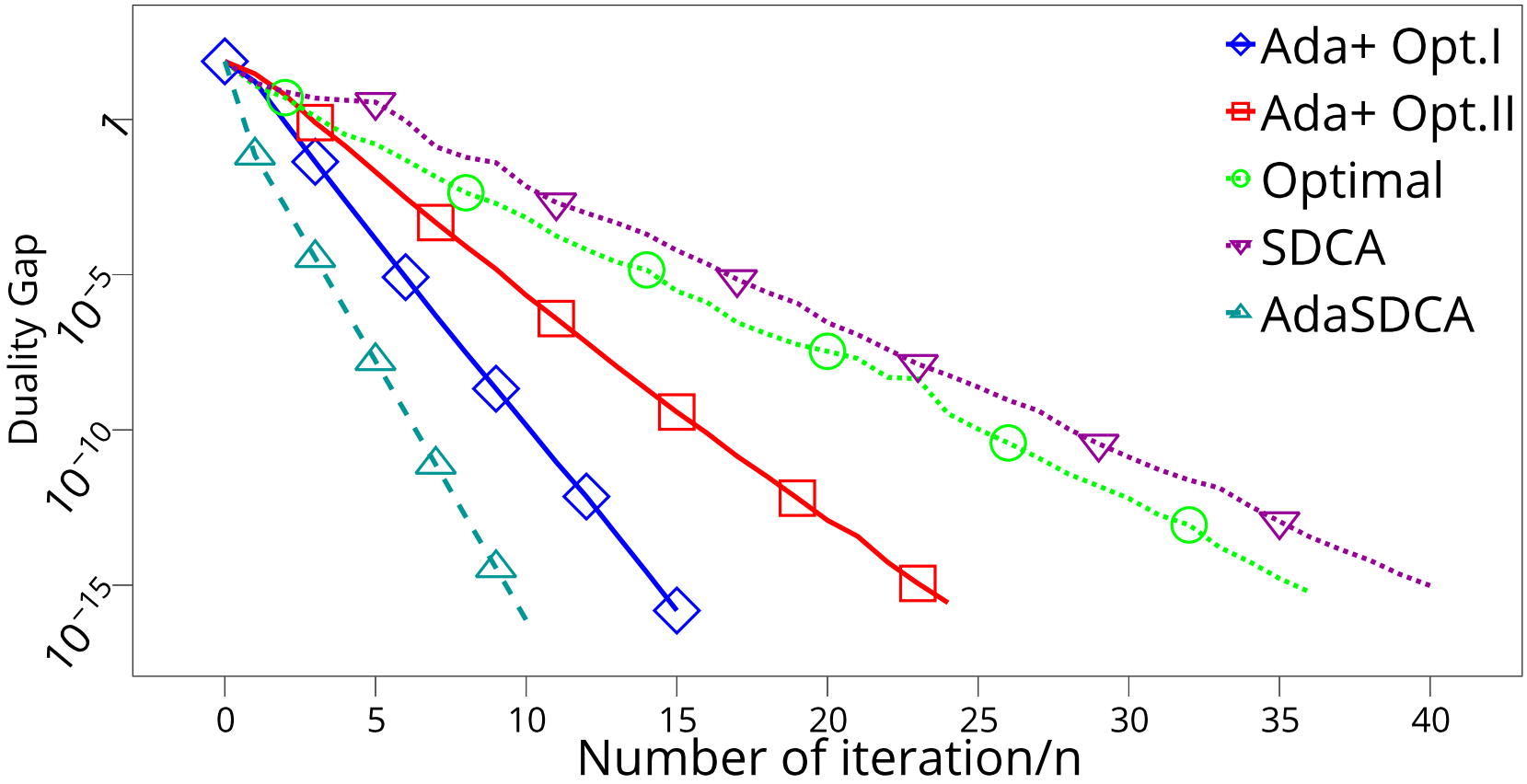

Figure 1: **w8a** dataset $d = 300$, $n = 49749$, Quadratic loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

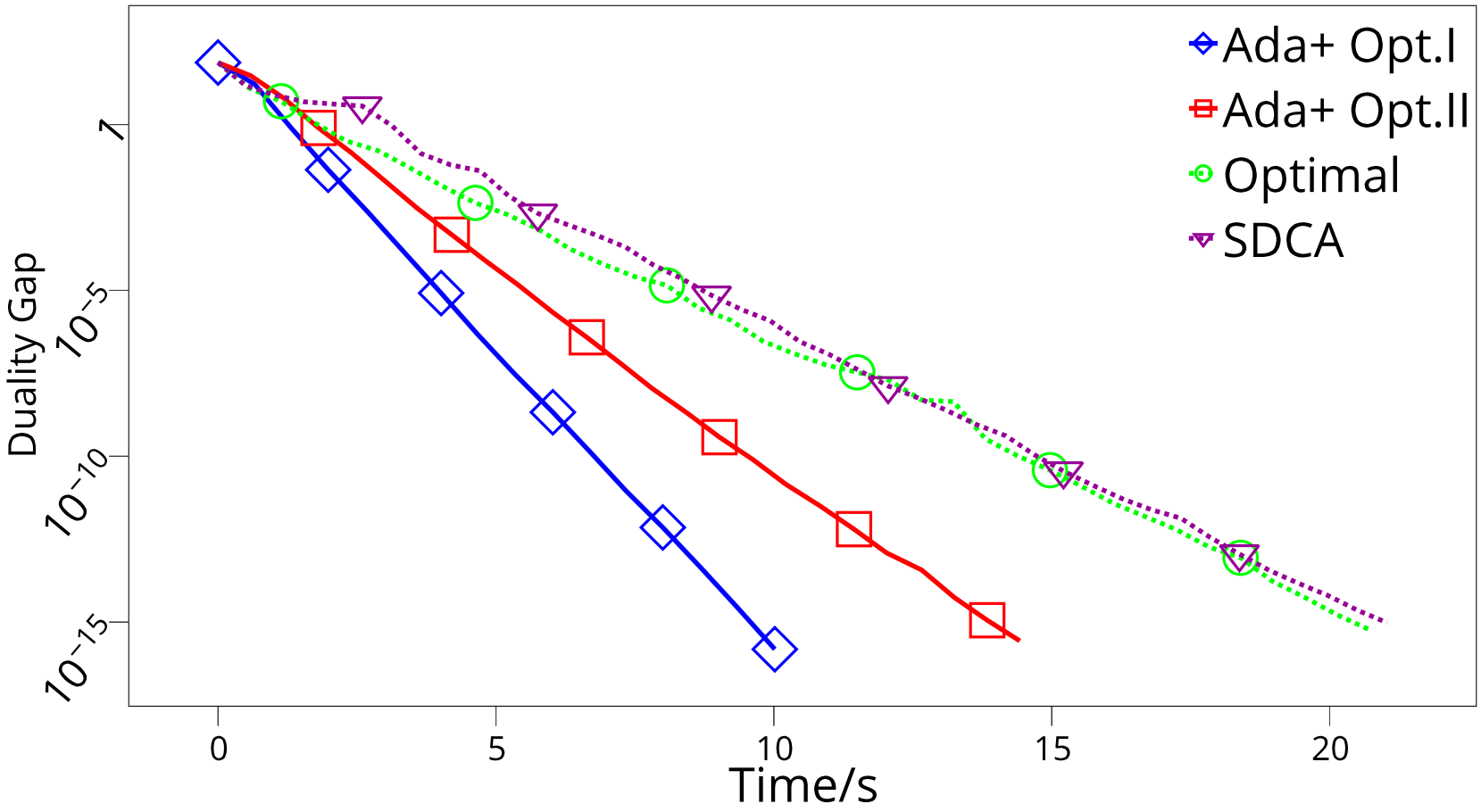

Figure 2: **w8a** dataset $d = 300$, $n = 49749$, Quadratic loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

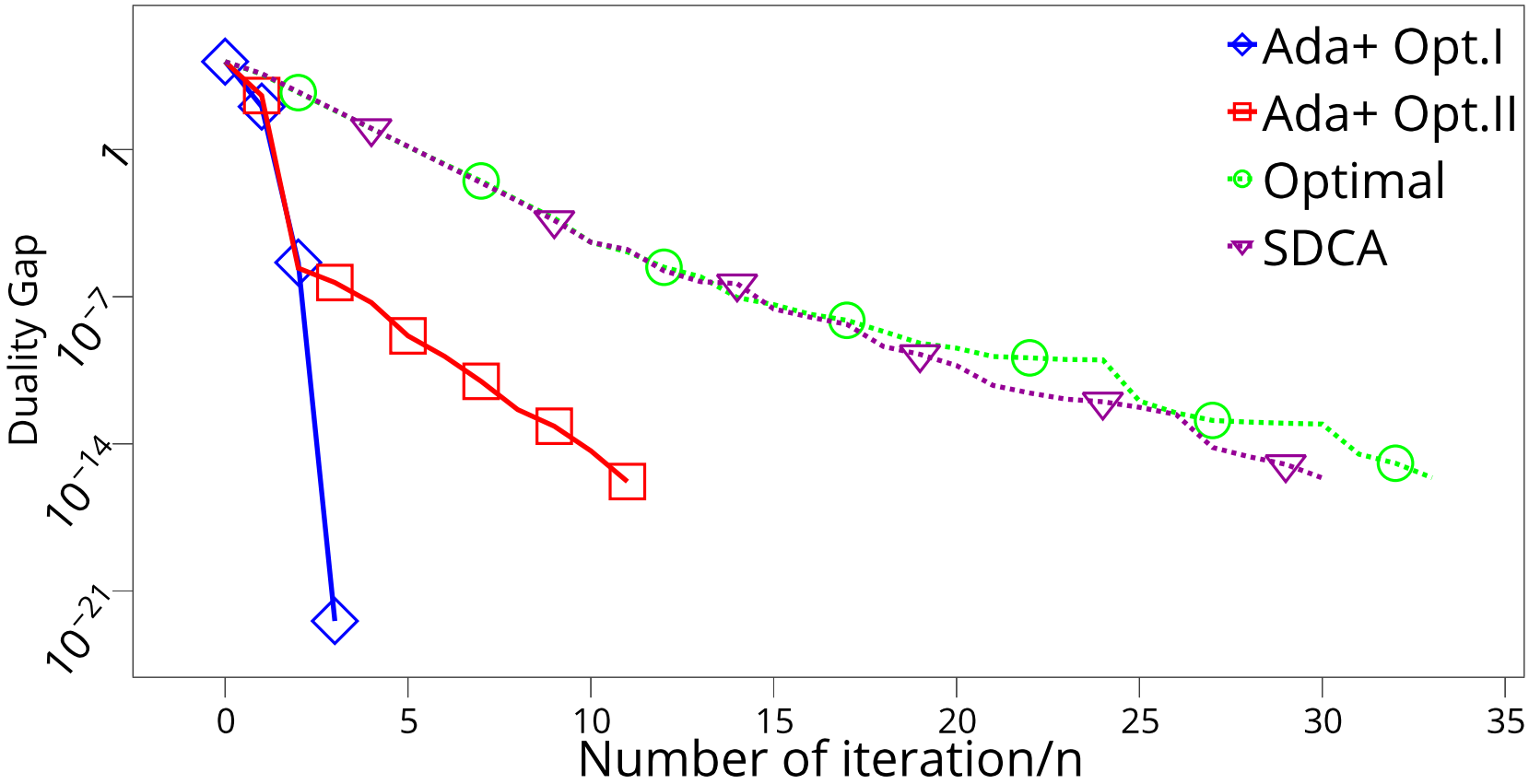

Figure 3: **cov1** dataset: $d = 54, n = 581,012$. Smooth Hinge loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

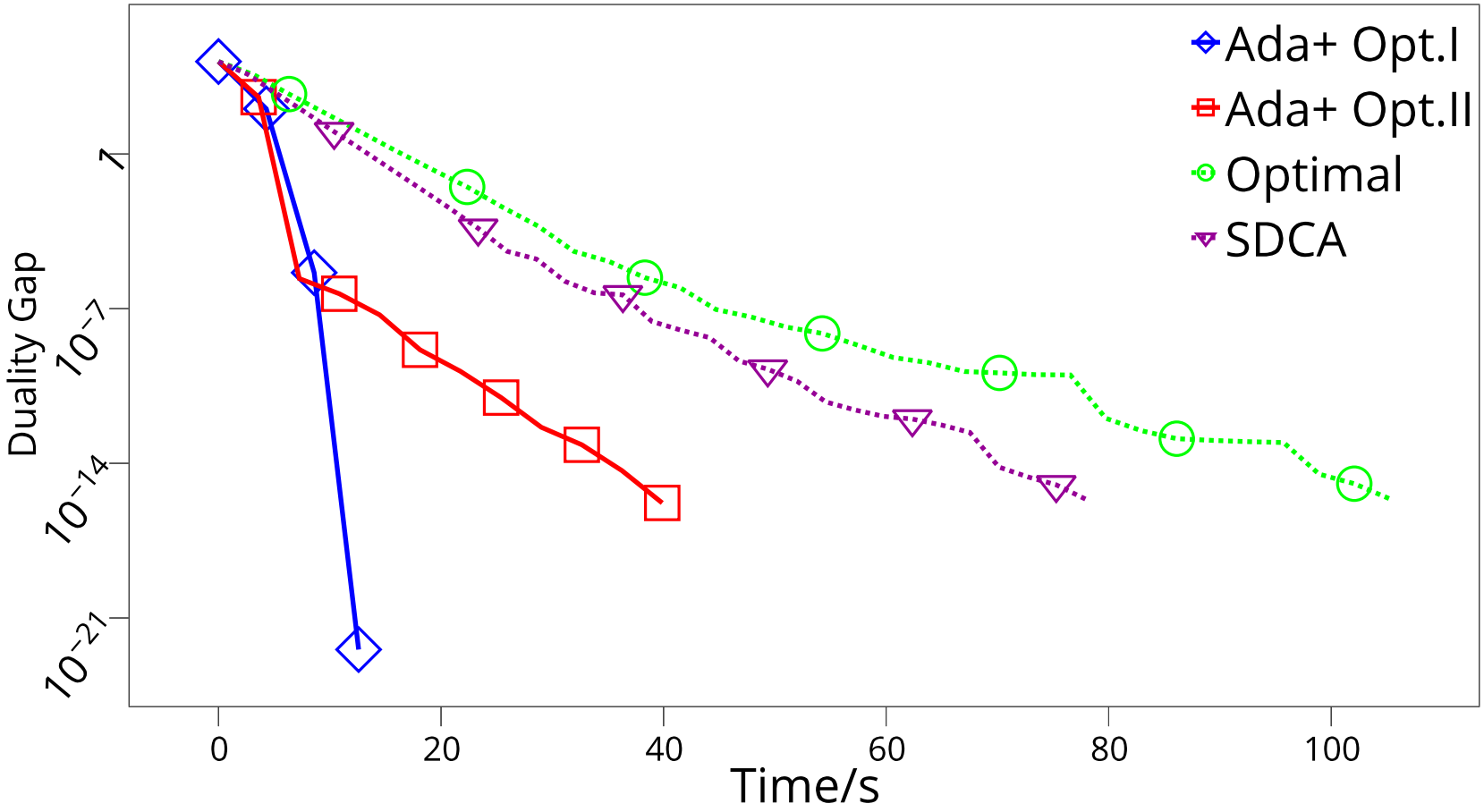

Figure 4: **cov1** dataset: $d = 54, n = 581,012$. Smooth Hinge loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

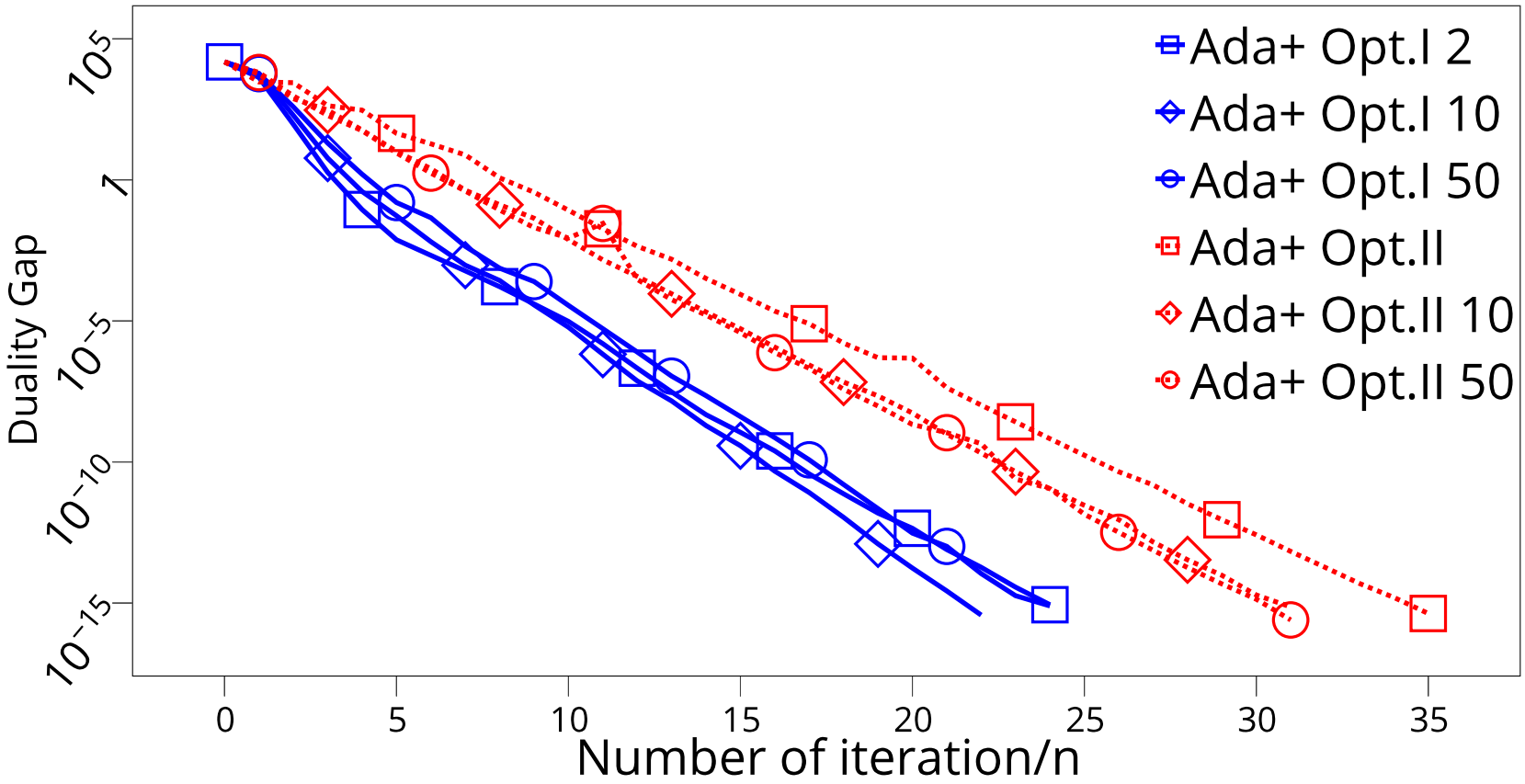

Figure 5: **cov1** dataset: $d = 54, n = 581,012$. Smooth Hinge loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$. comparison of different choices of the constant m.

Conclusion

1. This paper
   - AdaSDCA: an adaptive variant of stochastic dual coordinate ascent for ERM with
     - better complexity bound than the best fixed probability distribution
   - AdaSDCA+: a heuristic variant of AdaSDCA with
     - better numerical convergence speed
2. Future work
   - Dual free adaptive sampling probability
   - Adaptive sampling+SVRG/SAGA
   - Non-strongly convex case