Stochastic Dual Coordinate Ascent with Adaptive Probabilities: Supplementary material

Proofs

We shall need the following inequality.

Lemma 1. *Function $f:\mathbb{R}^n\to\mathbb{R}$ defined in (3) satisfies the following inequality:*

$$f(\alpha+h)\leq f(\alpha)+\langle\nabla f(\alpha),h\rangle+\frac{1}{2\lambda n^2}h^\top A^\top Ah, \quad (1)$$

holds for $\forall\alpha,h\in\mathbb{R}^n$.

Proof. Since g is 1-strongly convex, g^* is 1-smooth. Pick $\alpha,h\in\mathbb{R}^n$. Since, $f(\alpha)=\lambda g^*(\frac{1}{\lambda n}A\alpha)$, we have

$$\begin{aligned}
&f(\alpha+h)=\lambda g^*(\frac{1}{\lambda n}A\alpha+\frac{1}{\lambda n}Ah)\\
&\leq\lambda\left(g^*(\frac{1}{\lambda n}A\alpha)+\langle\nabla g^*(\frac{1}{\lambda n}A\alpha),\frac{1}{\lambda n}Ah\rangle+\frac{1}{2}\|\frac{1}{\lambda n}Ah\|^2\right)\\
&=f(\alpha)+\langle\nabla f(\alpha),h\rangle+\frac{1}{2\lambda n^2}h^\top A^\top Ah.
\end{aligned}$$

□

Proof of Lemma 3. It can be easily checked that the following relations hold

$$\nabla_i f(\alpha^t)=\frac{1}{n}A_i^\top w^t,\ \ \forall t\geq 0,\ \ i\in[n], \quad (2)$$

$$g(w^t)+g^*(\bar\alpha^t)=\langle w^t,\bar\alpha^t\rangle,\ \ \forall t\geq 0, \quad (3)$$

where $\{w^t,\alpha^t,\bar\alpha^t\}_{t\geq 0}$ is the output sequence of Algorithm 1. Let $t\geq 0$ and $\theta\in[0,\min_i p_i^t]$. For each $i\in[n]$, since ϕ_i is $1/\gamma$-smooth, ϕ_i^* is γ-strongly convex and thus for arbitrary $s_i\in[0,1]$,

$$\begin{aligned}
&\phi_i^*(-\alpha_i^t+s_i\kappa_i^t)\\
&=\phi_i^*\left((1-s_i)(-\alpha_i^t)+s_i\nabla\phi_i(A_i^\top w^t)\right)\\
&\leq(1-s_i)\phi_i^*(-\alpha_i^t)+s_i\phi_i^*(\nabla\phi_i(A_i^\top w^t))\\
&\quad-\frac{\gamma s_i(1-s_i)|\kappa_i^t|^2}{2}.
\end{aligned} \quad (4)$$

We have:

$$\begin{aligned}
&f(\alpha^{t+1})-f(\alpha^t)\\
&\overset{(1)}{\leq}\langle\nabla f(\alpha^t),\alpha^{t+1}-\alpha^t\rangle\\
&\qquad+\frac{1}{2\lambda n^2}\langle\alpha^{t+1}-\alpha^t,A^\top A(\alpha^{t+1}-\alpha^t)\rangle\\
&=\nabla_i f(\alpha^t)\Delta\alpha_{i_t}^t+\frac{v_i}{2\lambda n^2}|\Delta\alpha_{i_t}^t|^2\\
&\overset{(2)}{=}\frac{1}{n}A_{i_t}^\top w^t\Delta\alpha_{i_t}^t+\frac{v_i}{2\lambda n^2}|\Delta\alpha_{i_t}^t|^2
\end{aligned} \quad (5)$$

Thus,

$$\begin{aligned}
&D(\alpha^{t+1})-D(\alpha^t)\\
&\overset{(5)}{\geq}-\frac{1}{n}A_{i_t}^\top w^t\Delta\alpha_{i_t}^t-\frac{v_{i_t}}{2\lambda n^2}|\Delta\alpha_{i_t}^t|^2+\frac{1}{n}\sum_{i=1}^n\phi_i^*(-\alpha_i^t)\\
&\qquad-\frac{1}{n}\sum_{i=1}^n\phi_i^*(-\alpha_i^{t+1})\\
&=-\frac{1}{n}A_{i_t}^\top w^t\Delta\alpha_{i_t}^t-\frac{v_{i_t}}{2\lambda n^2}|\Delta\alpha_{i_t}^t|^2+\frac{1}{n}\phi_{i_t}^*(-\alpha_{i_t}^t)\\
&\quad-\frac{1}{n}\phi_{i_t}^*(-(\alpha_{i_t}^t+\Delta\alpha_{i_t}^t))\\
&=\max_{\Delta\in\mathbb{R}}-\frac{1}{n}A_{i_t}^\top w^t\Delta-\frac{v_{i_t}}{2\lambda n^2}|\Delta|^2+\frac{1}{n}\phi_{i_t}^*(-\alpha_{i_t}^t)\\
&\qquad-\frac{1}{n}\phi_{i_t}^*(-(\alpha_{i_t}^t+\Delta)),
\end{aligned}$$

where the last equality follows from the definition of $\Delta\alpha_{i_t}^t$ in Algorithm 1. Then by letting $\Delta=-s_i\kappa_{i_t}^t$ for some arbitrary $s_i\in[0,1]$ we get:

$$\begin{aligned}
&D(\alpha^{t+1})-D(\alpha^t)\\
&\geq\frac{s_iA_{i_t}^\top w^t\kappa_{i_t}^t}{n}-\frac{s_i^2v_{i_t}|\kappa_{i_t}^t|^2}{2\lambda n^2}+\frac{1}{n}\phi_{i_t}^*(-\alpha_{i_t}^t)\\
&\qquad-\frac{1}{n}\phi_{i_t}^*(-\alpha_{i_t}^t+s_i\kappa_{i_t}^t)\\
&\overset{(4)}{\geq}\frac{s_i}{n}\left(\phi_{i_t}^*(-\alpha_{i_t}^t)-\phi_{i_t}^*(\nabla\phi_{i_t}(A_{i_t}^\top w^t))+A_{i_t}^\top w^t\kappa_{i_t}^t\right)\\
&\qquad-\frac{s_i^2v_{i_t}|\kappa_{i_t}^t|^2}{2\lambda n^2}+\frac{\gamma s_i(1-s_i)|\kappa_{i_t}^t|^2}{2n}.
\end{aligned}$$

By taking expectation with respect to i_t we get:

$$\begin{aligned}
&\mathbb{E}_t\left[D(\alpha^{t+1}) - D(\alpha^t)\right] \\
&\geq \sum_{i=1}^n \frac{p_i^t s_i}{n}\left[\phi_i^*(-\alpha_i^t) - \phi_i^*(\nabla\phi_i(A_i^\top w^t)) + A_i^\top w^t \kappa_i^t\right] \\
&\quad - \sum_{i=1}^n \frac{p_i^t s_i^2 |\kappa_i^t|^2 (v_i + \lambda\gamma n)}{2\lambda n^2} + \sum_{i=1}^n \frac{p_i^t \gamma s_i |\kappa_i^t|^2}{2n}. \qquad (6)
\end{aligned}$$

Set

$$s_i = \begin{cases} 0, & i \notin I_t \\ \theta/p_i^t, & i \in I_t \end{cases} \qquad (7)$$

Then $s_i \in [0,1]$ for each $i \in [n]$ and by plugging it into (6) we get:

$$\begin{aligned}
&\mathbb{E}_t\left[D(\alpha^{t+1}) - D(\alpha^t)\right] \\
&\geq \frac{\theta}{n}\sum_{i \in I_t}\left[\phi_i^*(-\alpha_i^t) - \phi_i^*(\nabla\phi_i(A_i^\top w^t)) + A_i^\top w^t \kappa_i^t\right] \\
&\quad - \frac{\theta}{2\lambda n^2}\sum_{i \in I_t}\left(\frac{\theta(v_i + n\lambda\gamma)}{p_i^t} - n\lambda\gamma\right)|\kappa_i^t|^2
\end{aligned}$$

Finally note that:

$$\begin{aligned}
&P(w^t) - D(\alpha^t) \\
&= \frac{1}{n}\sum_{i=1}^n\left[\phi_i(A_i^\top w^t) + \phi_i^*(-\alpha_i^t)\right] + \lambda\left(g(w^t) + g^*(\bar{\alpha}^t)\right) \\
&\overset{(3)}{=} \frac{1}{n}\sum_{i=1}^n\left[\phi_i^*(-\alpha_i^t) + \phi_i(A_i^\top w^t)\right] + \frac{1}{n}\langle w^t, A\alpha^t\rangle \\
&= \frac{1}{n}\sum_{i=1}^n\left[\phi_i^*(-\alpha_i^t) + A_i^\top w^t \nabla\phi_i(A_i^\top w^t)\right. \\
&\qquad\quad \left. - \phi_i^*(\nabla\phi_i(A_i^\top w^t)) + A_i^\top w^t \alpha_i^t\right] \\
&= \frac{1}{n}\sum_{i=1}^n\left[\phi_i^*(-\alpha_i^t) - \phi_i^*(\nabla\phi_i(A_i^\top w^t)) + A_i^\top w^t \kappa_i^t\right] \\
&= \frac{1}{n}\sum_{i \in I_t}\left[\phi_i^*(-\alpha_i^t) - \phi_i^*(\nabla\phi_i(A_i^\top w^t)) + A_i^\top w^t \kappa_i^t\right]
\end{aligned}$$

□

Proof of Lemma 4. Note that (13) is a standard constrained maximization problem, where everything independent of p can be treated as a constant. We define the Lagrangian

$$L(p,\eta) = \theta(\kappa, p) - \eta\left(\sum_{i=1}^n p_i - 1\right)$$

and get the following optimality conditions:

$$\begin{aligned}
&\frac{|\kappa_i^t|^2 (v_i + n\lambda\gamma)}{p_i^2} = \frac{|\kappa_j^t|^2 (v_j + n\lambda\gamma)}{p_j^2}, \ \forall i, j \in [n] \\
&\sum_{i=1}^n p_i = 1 \\
&p_i \geq 0, \ \ \forall i \in [n],
\end{aligned}$$

the solution of which is (14). □

Proof of Lemma 5. Note that in the proof of Lemma 3, the condition $\theta \in [0, \min_{i \in I_t} p_i^t]$ is only needed to ensure that s_i defined by (7) is in $[0,1]$ so that (4) holds. If ϕ_i is quadratic function, then (4) holds for arbitrary $s_i \in \mathbb{R}$. Therefore in this case we only need θ to be positive and the same reasoning holds. □

Additional Numerical Experiments

We now provide more numerical experiments.

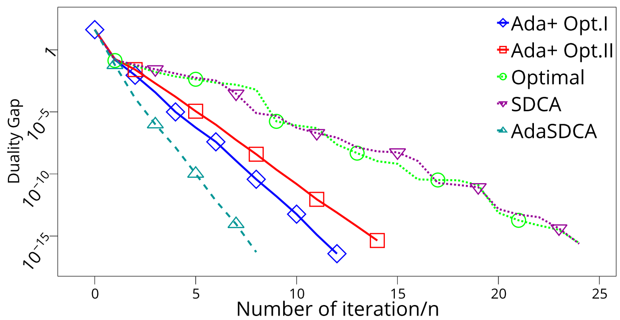

Figure 10. dorothea dataset d = 100000, n = 800, Quadratic loss with L_2 regularizer, comparing number of iterations with known algorithms

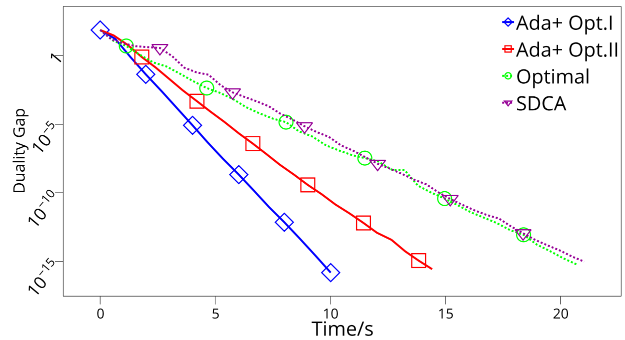

Figure 13. w8a dataset d = 300, n = 49749, Quadratic loss with L_2 regularizer, comparing real time with known algorithms

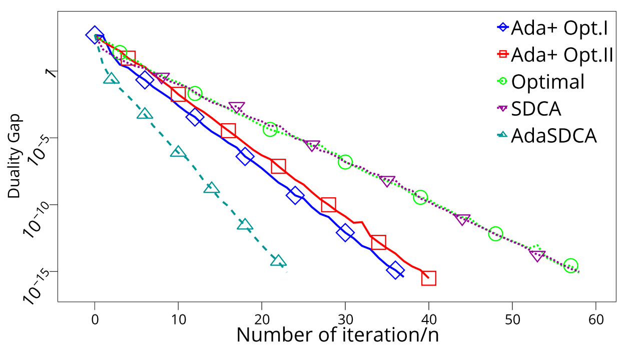

Figure 11. mushrooms dataset d = 112, n = 8124, Quadratic loss with L_2 regularizer, comparing number of iterations with known algorithms

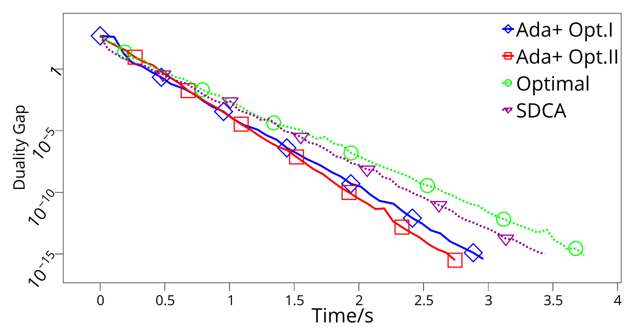

Figure 14. mushrooms dataset d = 112, n = 8124, Quadratic loss with L_2 regularizer, comparing real time with known algorithms

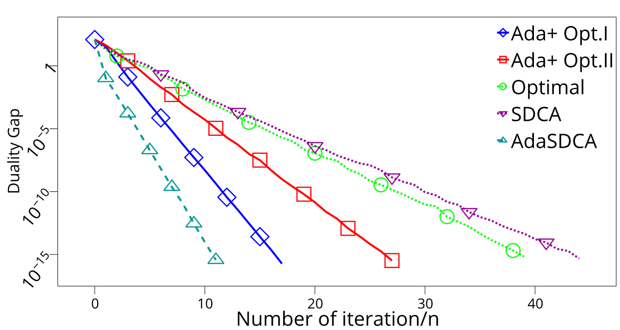

Figure 12. ijcnn1 dataset d = 22, n = 49990, Quadratic loss with L_2 regularizer, comparing number of iterations with known algorithms

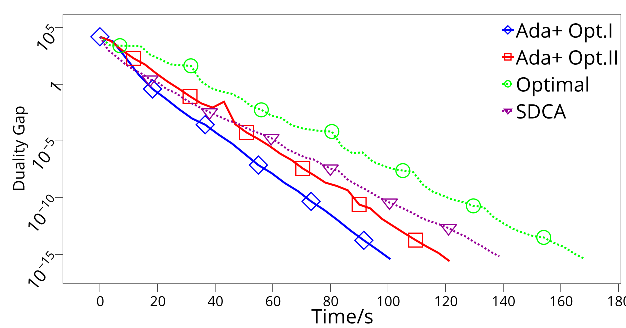

Figure 15. cov1 dataset d = 54, n = 581012, Quadratic loss with L_2 regularizer, comparing real time with known algorithms

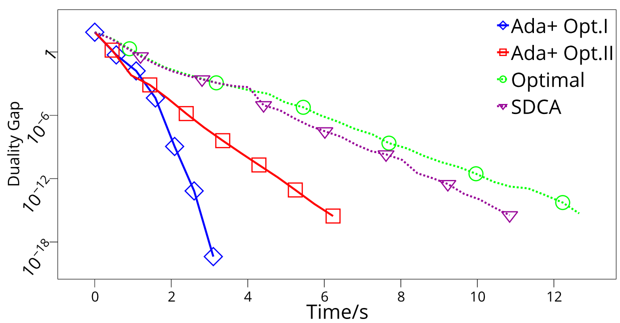

Figure 16. w8a dataset $d = 300$, $n = 49749$, Smooth Hinge loss with L_2 regularizer, comparing real time with known algorithms

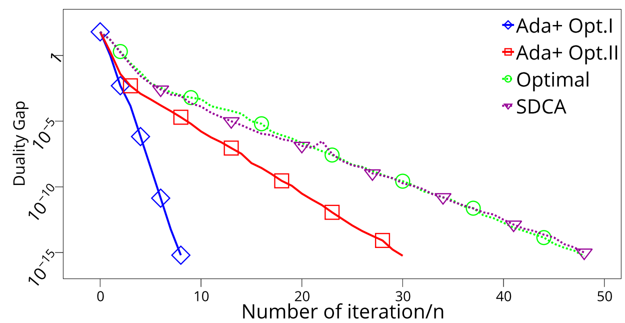

Figure 19. mushrooms dataset $d = 112$, $n = 8124$, Smooth Hinge loss with L_2 regularizer, comparing number of iterations with known algorithms

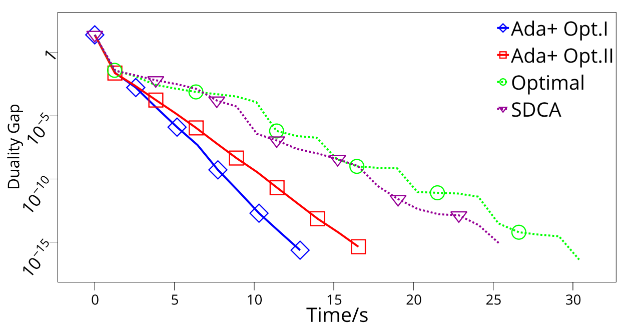

Figure 17. dorothea dataset $d = 100000$, $n = 800$, Smooth Hinge loss with L_2 regularizer, comparing real time with known algorithms

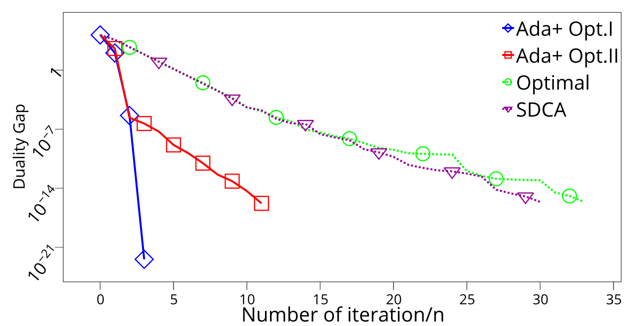

Figure 20. cov1 dataset $d = 54$, $n = 581012$, Smooth Hinge loss with L_2 regularizer, comparing number of iterations with known algorithms

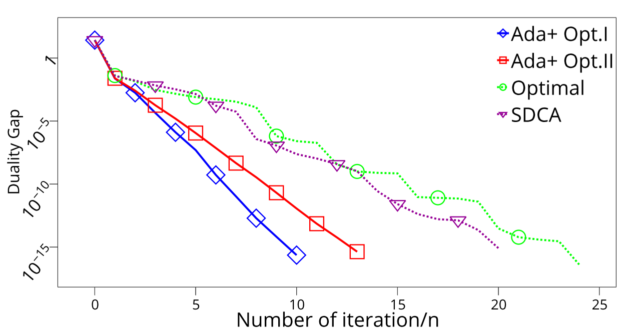

Figure 18. dorothea dataset $d = 100000$, $n = 800$, Smooth Hinge loss with L_2 regularizer, comparing number of iterations with known algorithms

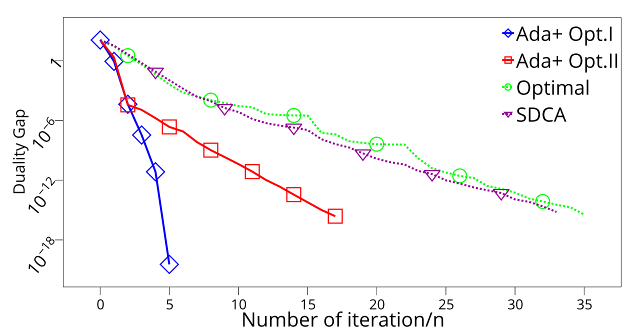

Figure 21. ijcnn1 dataset $d = 22$, $n = 49990$, Smooth Hinge loss with L_2 regularizer, comparing number of iterations with known algorithms

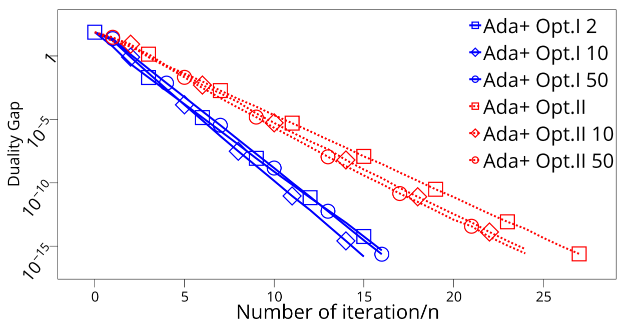

Figure 22. w8a dataset d = 300, n = 49749, Quadratic loss with L_2 regularizer, comparison of different choices of the constant m

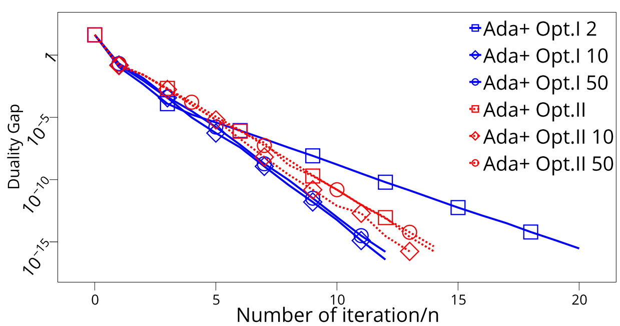

Figure 23. dorothea dataset d = 100000, n = 800, Quadratic loss with L_2 regularizer, comparison of different choices of the constant m

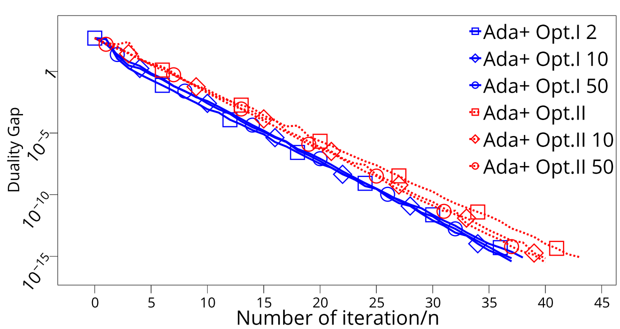

Figure 24. mushrooms dataset d = 112, n = 8124, Quadratic loss with L_2 regularizer, comparison of different choices of the constant m

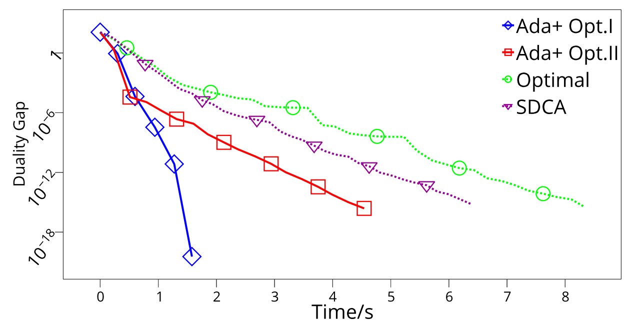

Figure 25. ijcnn1 dataset d = 22, n = 49990, Smooth Hinge loss with L_2 regularizer, comparing real time with known algorithms

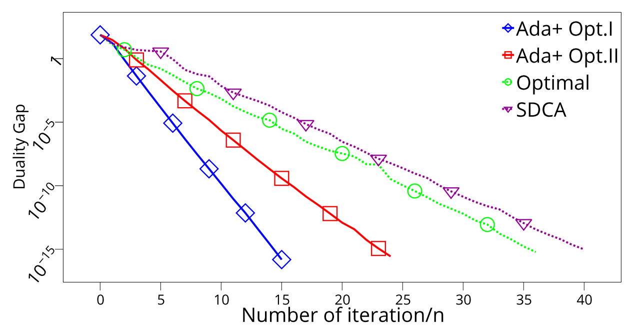

Figure 26. w8a dataset d = 300, n = 49749, Quadratic loss with L_2 regularizer, comparing number of iterations with known algorithms

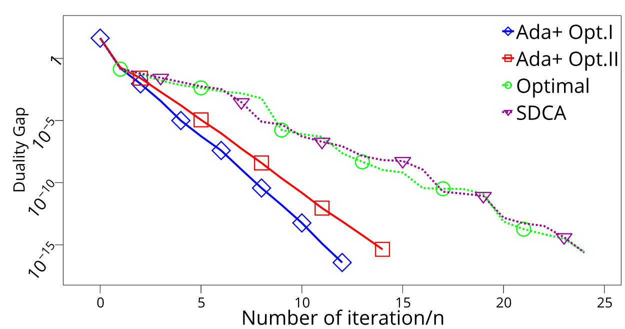

Figure 27. dorothea dataset d = 100000, n = 800, Quadratic loss with L_2 regularizer, comparing number of iterations with known algorithms

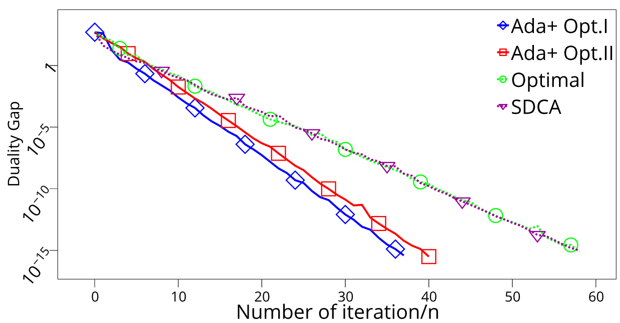

Figure 28. mushrooms dataset d = 112, n = 8124, Quadratic loss with L_2 regularizer, comparing number of iterations with known algorithms

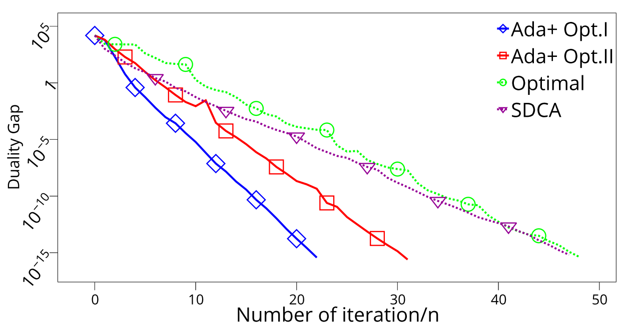

Figure 29. cov1 dataset d = 54, n = 581012, Quadratic loss with L_2 regularizer, comparing number of iterations with known algorithms

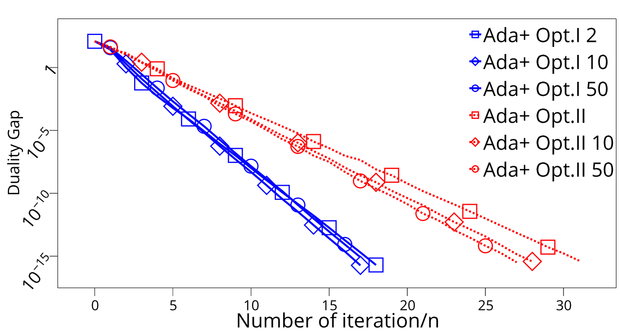

Figure 30. ijcnn1 dataset d = 22, n = 49990, Quadratic loss with L_2 regularizer, comparison of different choices of the constant m

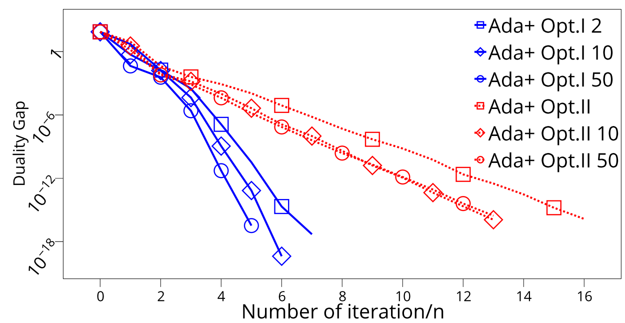

Figure 31. w8a dataset d = 300, n = 49749, Smooth Hinge loss with L_2 regularizer, comparison of different choices of the constant m

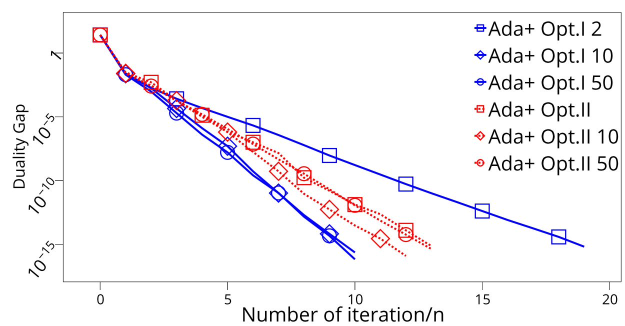

Figure 32. dorothea dataset d = 100000, n = 800, Smooth Hinge loss with L_2 regularizer, comparison of different choices of the constant m

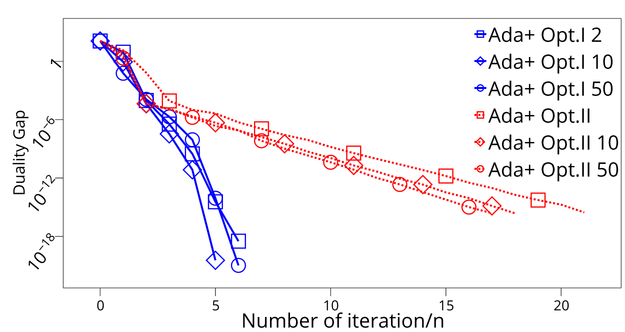

Figure 33. ijcnn1 dataset d = 22, n = 49990, Smooth Hinge loss with L_2 regularizer, comparison of different choices of the constant m