A Proofs

Proof of Lemma 1

Since $J_j(w) \leq J_t(w) \leq J(w)$ for all $j \leq i$ this property also applies to their respective minima. Moreover, since w^* minimizes $J(w)$ we have $J(w^*) \leq J(w_t)$. Since Taylor expansions are exact at the point of expansion $J(w_t) = J_{t+1}(w_t)$. The first inequality follows from the definition of ϵ_t, and the fact that J_t is monotonically increasing. Finally, since $J_{j+1}(w_j) = J(w_j)$ it is easy to see that $\gamma_t \geq \epsilon_t \geq \min_{t' \leq t} J(w_{t'}) - J_t(w_t) \geq \min_{t' \leq t} J(w_{t'}) - J(w^*)$.

Proof of Theorem 2

We rewrite (5) as a constrained optimization problem: $\text{minimize}\, \lambda\Omega(w) + \xi$ subject to $\xi \geq 0$ and $\xi \geq \langle a_{t'}, w\rangle + b_{t'}$ for all $t' \leq t$. The corresponding Lagrange function can be written as

$$L(w, \xi, \alpha) = \lambda\Omega(w) + \xi - \alpha^\top \left(\xi\mathbf{e} - A^\top w - b_t\right) \text{ with } \alpha \geq 0, \tag{16}$$

where $\mathbf{e}$ is the vector of all ones, and $\alpha \geq 0$ denotes that each component of α is non-negative. Taking derivatives with respect to ξ yields $1 - \alpha^\top \mathbf{e} = 0$. Moreover, minimization of L with respect to w implies solving $\text{maximize}_w \left\langle w, -\lambda^{-1} A\alpha\right\rangle - \Omega(w) = \Omega^*(-\lambda^{-1}A\alpha)$. Plugging both terms back into (16) allows us to eliminate the primal variables ξ and w.

Proof of Theorem 4

To show Theorem 4 we need several technical intermediate steps. The first one allows us to bound the maximum value of a concave function provided that we know its first derivative and a bound on the second derivative.

Lemma 7 *Denote by $f : [0, 1] \to \mathbb{R}$ a concave function with $f(0) = 0$, $f'(0) = l$, and $|f''(x)| \leq H$. Then for all $x \in [0, 1]$ we have $\max_{x\in[0,1]} f(x) \geq \frac{l}{2}\min(1, l/H)$.*

Proof We first observe that $lx - \frac{H}{2}x^2 \leq f(x)$. The unconstrained maximum of the lower bound is attained by l/H leading to a maximum value of $l^2/2H$. If $l > H$ we pick $x = 1$ which yields $l - H/2$. The latter is majorized by $l/2$. Taking the minimum over both maxima proves the claim. ■

To apply the above result, we need to compute the gradient and Hessian of $J^*_{t+1}(\alpha)$ with respect to the line search direction $((1-\eta)\alpha^t, \eta)$. The following lemma takes care of the gradient:

Lemma 8 *Denote by α_t the solution of (5) at time instance t. Moreover, denote by $\bar{A} = [A, a_{t+1}]$ and $\bar{b} = [b, b_{t+1}]$ the extended matrices and vectors needed to define the dual problem for step $t + 1$, and let $\bar{\alpha} \in \mathbb{R}^{t+1}$. Then the following holds:*

$$\partial_{\bar{\alpha}} J^*_{t+1}([\alpha_t, 0]) = \bar{A}^\top w_t + \bar{b} \text{ and} \tag{17}$$

$$[-\alpha_t, 1]^\top \left[\bar{A}^\top w_t + \bar{b}\right] = J_{t+1}(w_t) - J_t(w_t) = \gamma_t \tag{18}$$

Proof By the dual connection $\partial\Omega^*(-\lambda^{-1}A\alpha_t) = w_t$. Hence we have that $\partial_{\bar{\alpha}} - \lambda\Omega^*(-\lambda^{-1}\bar{A}\bar{\alpha}) + \bar{\alpha}^\top\bar{b} = \bar{A}^\top w_t + \bar{b}$ for $\bar{\alpha} = [\alpha_t, 0]^\top$. This proves the first claim. To see the second part we eliminate ξ from of the Lagrangian (16) and write the partial Lagrangian

$$L(w, \alpha) = \lambda\Omega(w) + \alpha^\top \left(A^\top w + b\right) \text{ with } \alpha \geq 0. \tag{19}$$

The result follows by noting that at optimality $L(w_t, \alpha_t) = J_t(w_t)$ and $J_{t+1}(w_t) = \lambda\Omega(w_t) + \langle a_{t+1}, w_t\rangle + b_{t+1}$. Consequently we have

$$J_{t+1}(w_t) - J_t(w_t) = \lambda\Omega(w_t) + \langle a_{t+1}, w_t\rangle + b_t - \lambda\Omega(w_t) - \alpha^t(A^\top w_t + b)$$

Rearranging terms proves the claim. ■

To apply Lemma 7 we also need to bound the second derivative.

Lemma 9 *Under the assumptions of Lemma 8 we have*

$$\partial^2_{\bar{\alpha}} J^*_{t+1}(\bar{\alpha}) = -\lambda^{-1}\bar{A}^\top \partial^2\Omega^*(-\lambda^{-1}\bar{A}\bar{\alpha})\bar{A} \tag{20}$$

$$\textit{moreover } \bar{A}[-\alpha_t, 1] = \partial_w J(w_t) \tag{21}$$

Proof The first equality follows by the chain rule. Next note that $\partial_w \Omega(w_t) = -\lambda^{-1} A\alpha_t$ by dual connection. Since $a_{t+1} = \partial_w R(w_t)$ the claim follows from $J(w) = R_{\mathrm{emp}}(w) + \lambda\Omega(w)$. ■

This result allows us to express the second derivative of the dual objective function (6) in terms of the gradient of the risk functional. The idea is that as we approach optimality, the second derivative will vanish. We will use this fact to argue that for continuously differentiable losses $R_{\mathrm{emp}}(w)$ we enjoy linear convergence throughout.

Proof [Theorem 4] We overload the notation for J^*_{t+1} by defining the following one dimensional concave function

$$J^*_{t+1}(\eta) := J^*_{t+1}(((1-\eta)\alpha^t, \eta)) = -\lambda\Omega^*(-\lambda^{-1}\bar{A}[(1-\eta)\alpha_t^\top, \eta]) + [(1-\eta)\alpha_t^\top, \eta]\bar{b}. \tag{22}$$

Clearly, $J^*_{t+1}(0) = J_t(w_t)$. Furthermore, by (18), (20), and (21) it follows that

$$\partial_\eta J^*_{t+1}(\eta)|_{\eta=0} = [-\alpha_t, 1]^\top \partial_{\bar{\alpha}} J^*_{t+1}([\alpha_t, 0]) = \gamma_t \text{ and} \tag{23}$$

$$\partial^2_\eta J^*_{t+1}(\eta) = -\lambda^{-1}[-\alpha_t, 1]^\top \bar{A}^\top \partial^2\Omega^*(-\lambda^{-1}\bar{A}[(1-\eta)\alpha_t, \eta])\bar{A}[-\alpha_t, 1]^\top \tag{24}$$

$$= -\lambda^{-1}[\partial_w J(w_t)]^\top \partial^2\Omega^*(-\lambda^{-1}\bar{A}[(1-\eta)\alpha_t, \eta])[\partial_w J(w_t)] := q. \tag{25}$$

By our assumption on $\|\partial^2\Omega^*\| \leq H^*$ we have

$$|q| \leq H^* \|\partial_w J(w_t)\|^2/\lambda.$$

Next we need to bound the gradient of J. For this purpose note that $\partial_w \lambda\Omega(w_t) = -A^\top \alpha_t$ and moreover that $\|\alpha_t\| = 1$. This implies that $\partial_w \lambda\Omega(w_t)$ lies in the convex hull of the past gradients, $a_{t'}$. By our assumption on $\|\partial_w R_{\mathrm{emp}}(w)\|$ it follows that $\|\partial_w \lambda\Omega(w_t)\| \leq G$. We conclude that

$$\|\partial_w J(w_t)\|^2 \leq 4G^2 \text{ and } q \leq 4G^2 H^*/\lambda.$$

Invoking Lemma 7 on $J^*_{t+1}(\eta) - J_t(w_t)$ proves the first part of (11). Since $\gamma_t \geq \epsilon_t$ the second part of the inequality follows.

For the second part note that (11) already yields the $\gamma_t/2$ decrease when $\gamma_t \geq 4G^2H^*/\lambda$. To show the other parts we need to show that the gradient of the regularized risk vanishes as we converge to the optimal solution. Towards this end, we apply Lemma 7 in the *primal*. This allows us to bound $\|\partial_w J(w_t)\|$ in terms of γ_t. Plugging in the first and second derivative of $J(w_t)$ we obtain

$$\gamma_t \geq \frac{1}{2}\|\partial_w J(w_t)\| \min(1, \|\partial_w J(w_t)\|/H).$$

If $\|\partial_w J(w_t)\| > H$, then $\gamma_t \geq \frac{1}{2}\|\partial_w J(w_t)\|$ which in turn yields $|q| \leq 4\lambda^{-1}\gamma_t^2 H^*$. Plugging this into Lemma 9 yields a lower bound on the improvement of $\lambda/8H^*$.

Finally, for $\|\partial_w J(w_t)\| \leq H$ we have $\gamma_t \geq \|\partial_w J(w_t)\|^2/2H$, which implies $|q| \leq 2HH^*\gamma_t/\lambda$. Plugging this into Lemma 7 yields an improvement of $\lambda\gamma_t/4HH^*$.

Since both cases cover the remaining range of convergence, the minimum $\min(\lambda/8H^*, \lambda\gamma_t/4HH^*)$ provides a lower bound for the improvement. The crossover point between both terms occurs at $\gamma_t = H/2$. Rearranging the conditions leads to the (pessimistic) improvement guarantees of the second claim. ■

Proof of Theorem 5

For any $\epsilon_t > 4G^2H^*/\lambda$ it follows from (11) that $\epsilon_{t+1} \leq \epsilon_t/2$. Moreover, $\epsilon_0 \leq J(0)$, since we know that J is nonnegative. Hence we need at most $\log_2[\lambda J(0)/4G^2H^*]$ to achieve this level of precision. Subsequently we need to solve the following difference equation

$$\epsilon_{t+1} - \epsilon_t = -\frac{\lambda}{8G^2H^*}\epsilon_t^2. \tag{26}$$

Since this is monotonically decreasing, we can upper bound this by solving the differential equation $\epsilon'(t) = -\epsilon^2(t)\lambda/8G^2H^*$, with the boundary conditions $\epsilon(0) = 4G^2H^*/\lambda$. This in turn yields $\epsilon(t) = \frac{8G^2H^*}{\lambda(t+2)}$, and hence $t \leq \frac{8G^2H^*}{\lambda\epsilon} - 2$. For a given ϵ we will need $\frac{8G^2H^*}{\lambda\epsilon} - 2$ more steps to converge. This proves the first claim.

To analyze convergence in the second case we need to study two additional phases: for $\epsilon_t \in [H/2, 4G^2H^*/\lambda]$ we see constant progress. Hence it takes us $4\lambda^{-2}[8G^2(H^*)^2 - HH^*\lambda]$ steps to cover this interval. Finally in the third phase we have $\epsilon_{t+1} \leq \epsilon_t[1 - \lambda/4HH^*]$. Starting from $\epsilon_t = H/2$ we need $\log_2[2\epsilon/H]/\log_2[1 - \lambda/4HH^*]$ steps to converge. Expanding the logarithm in the denominator close to 1 proves the claim.

Algorithm 2 Line Search

Initialize $w_0 = 0$, $r_0 = 0$, $t = 0$
repeat
  Compute gradient a_{t+1} and offset b_{t+1}
  Compute $\gamma_t = R_{\mathrm{emp}}(w_t) + \lambda \|w_t\|^2 - r_t$
  Compute η using Eq. (15)
  Update $w_{t+1} \leftarrow (1-\eta) w_t - (\eta/\lambda) a_{t+1}$
  Update $r_{t+1} \leftarrow (1-\eta) r_t - \eta r_{t+1}$
  $t \leftarrow t + 1$
until $\epsilon_t \leq \epsilon$

B The Linesearch algorithm

We describe the linesearch algorithm we used in Algorithm 2.

C Experiments

C.1 Datasets

We present results on the datasets listed in Table 1. Our choice is largely influenced by the choices in [4, 5] concerning classification tasks for the purpose of reproducing results reported in their work.

dataset	abbr.	#examples	dimension	density %
Adult9	adult9	32561	123	11.28
Real & Simulated	real-sim	57763	20958	0.25
Autos & Aviation	aut-avn	56862	20707	0.25
Web8	web8	33546	300	4.24
KDDCup-99	kdd99	3398431	127	12.86
Reuters CCAT	ccat	804414	47236	0.16
Reuters C11	c11	804414	47236	0.16
Arxiv astro-ph	astro-ph	62369	99757	0.08
Covertype	covertype	581012	54	22.22

Table 1: Datasets their properties.

C.2 Convergence

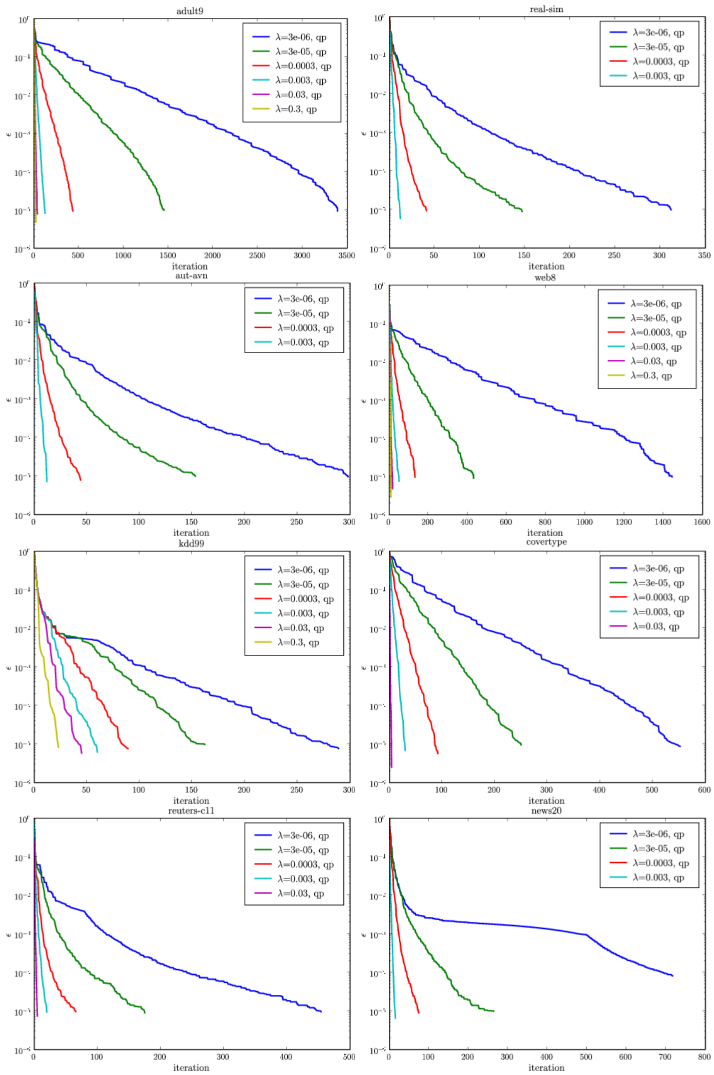

Figure 4: Convergence of the bundle method (semilog-y) with a QP inner loop. We plot ϵ_t as a function of number of iterations for various datasets. Note that there is a potential numerical error in the experiments reported for the news20 dataset.

C.3 Comparison with Pegasos

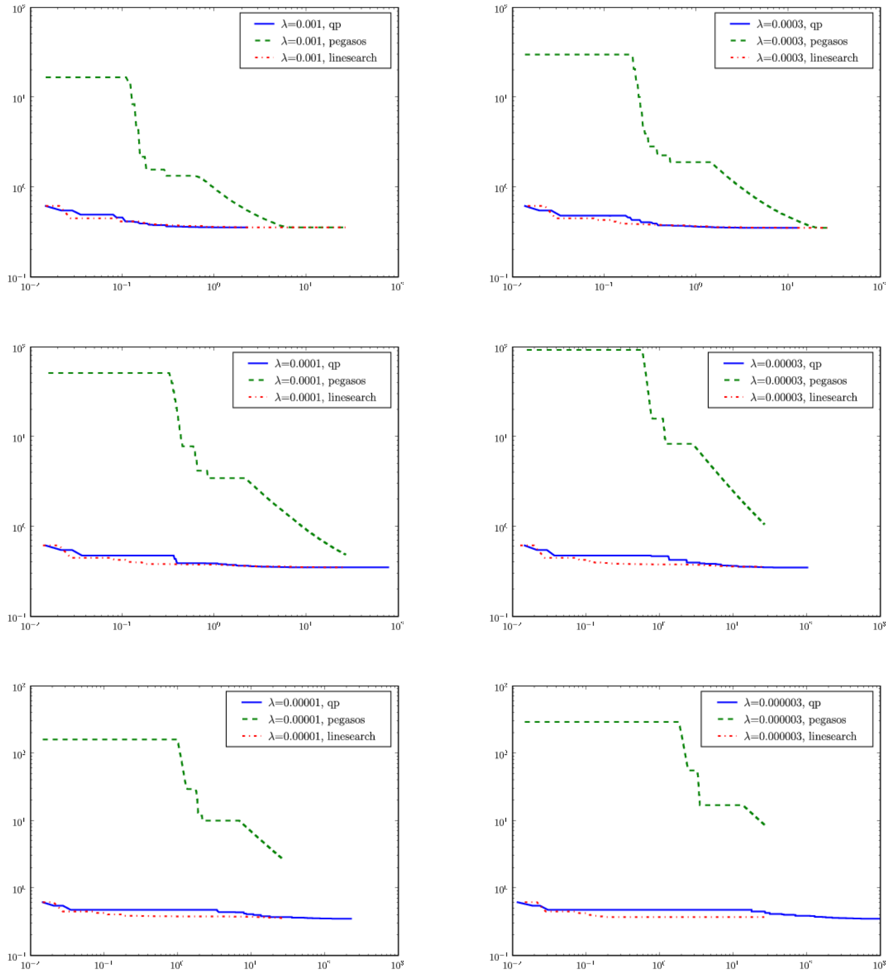

Figure 5: Objective function value with respect to time on adult9 with various values of λ.

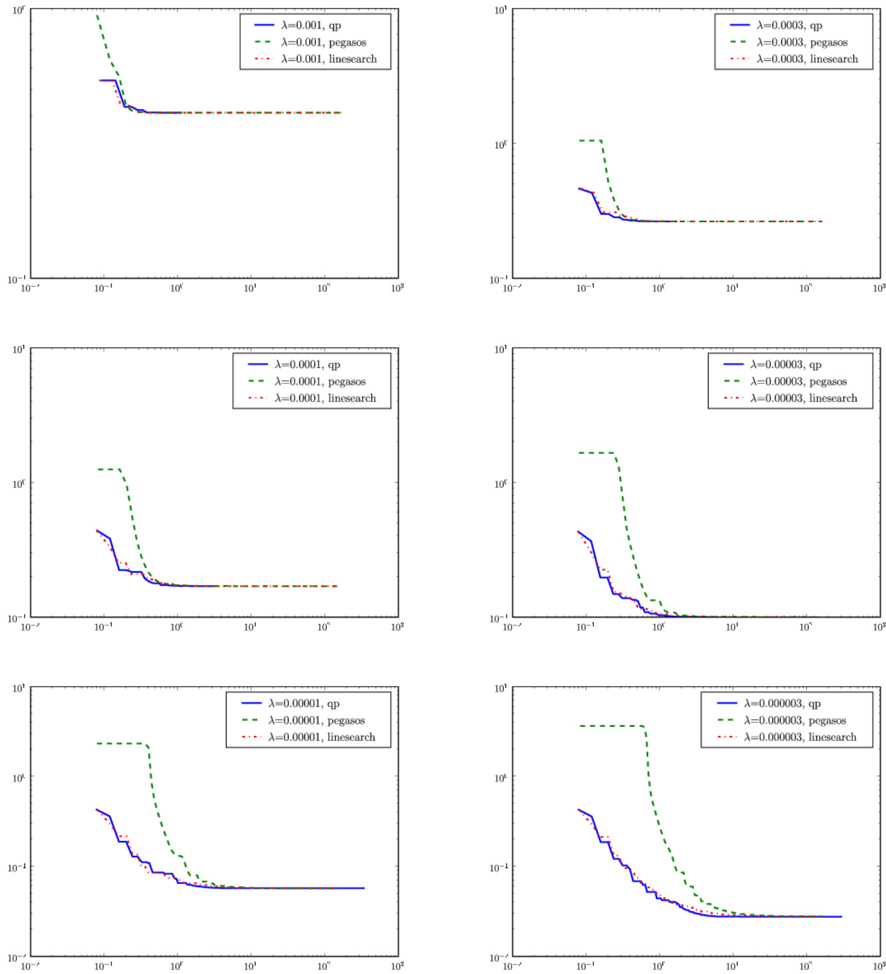

Figure 6: Objective function value with respect to time on real-sim with various values of λ.

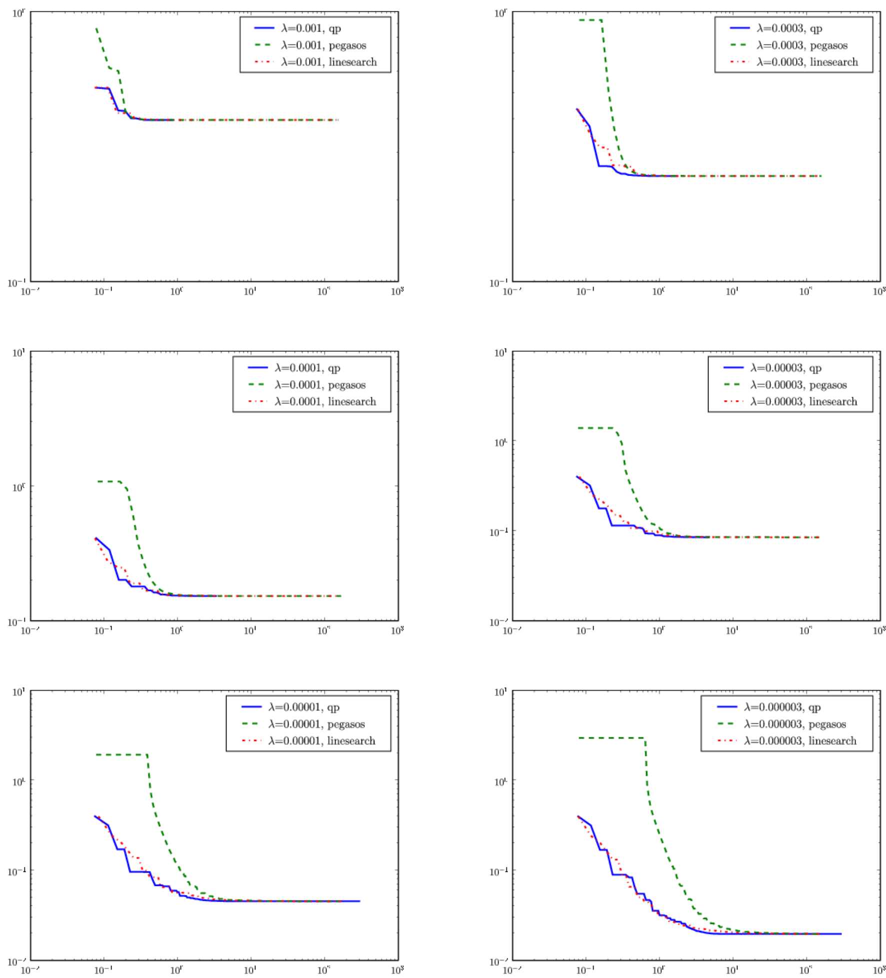

Figure 7: Objective function value with respect to time on aut-avn with various values of λ.

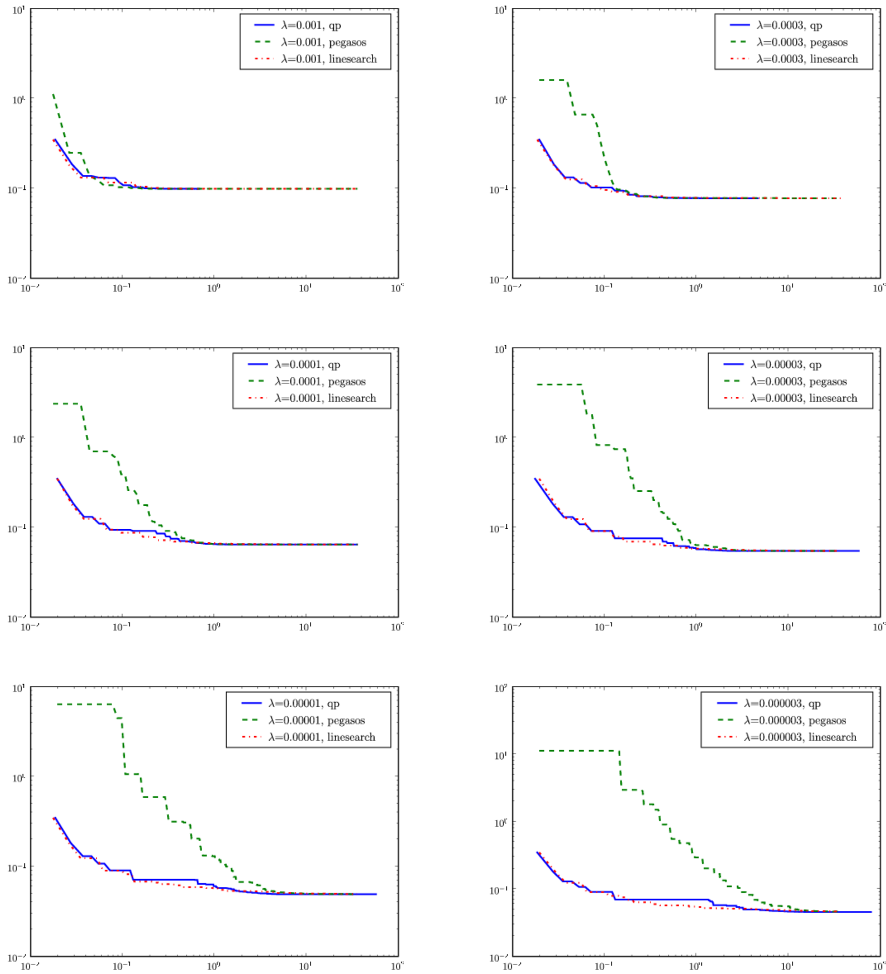

Figure 8: Objective function value with respect to time on web8 with various values of λ.

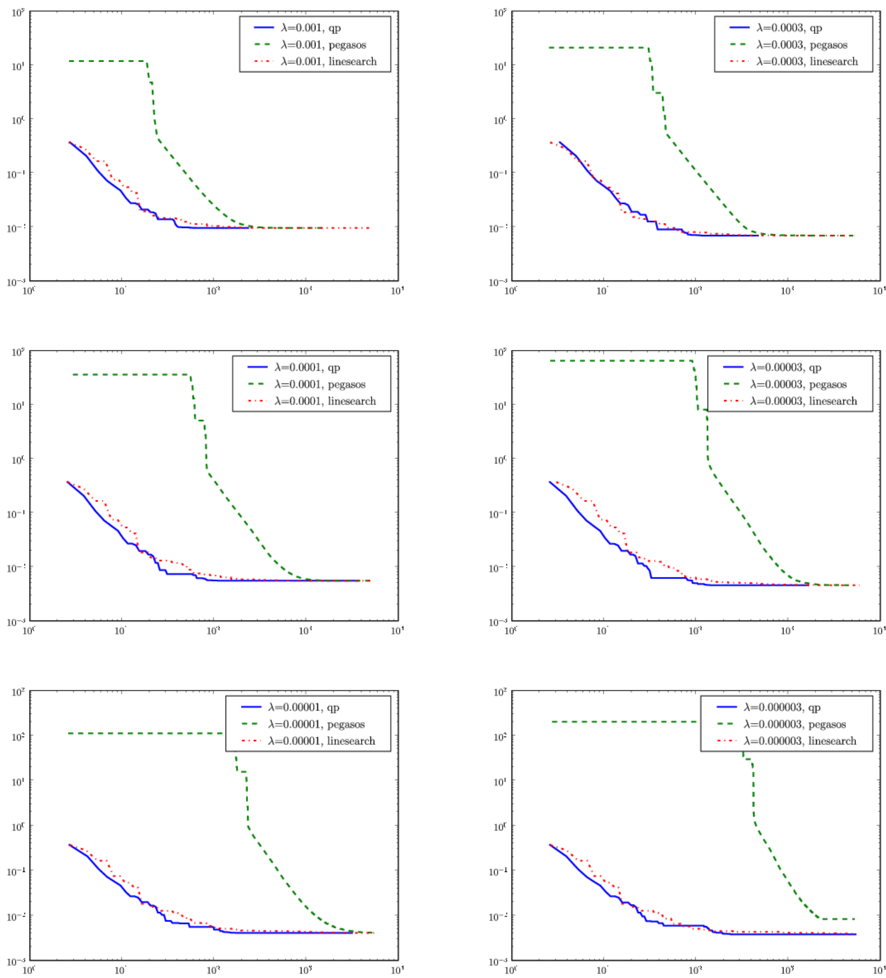

Figure 9: Objective function value with respect to time on kdd99 with various values of λ.

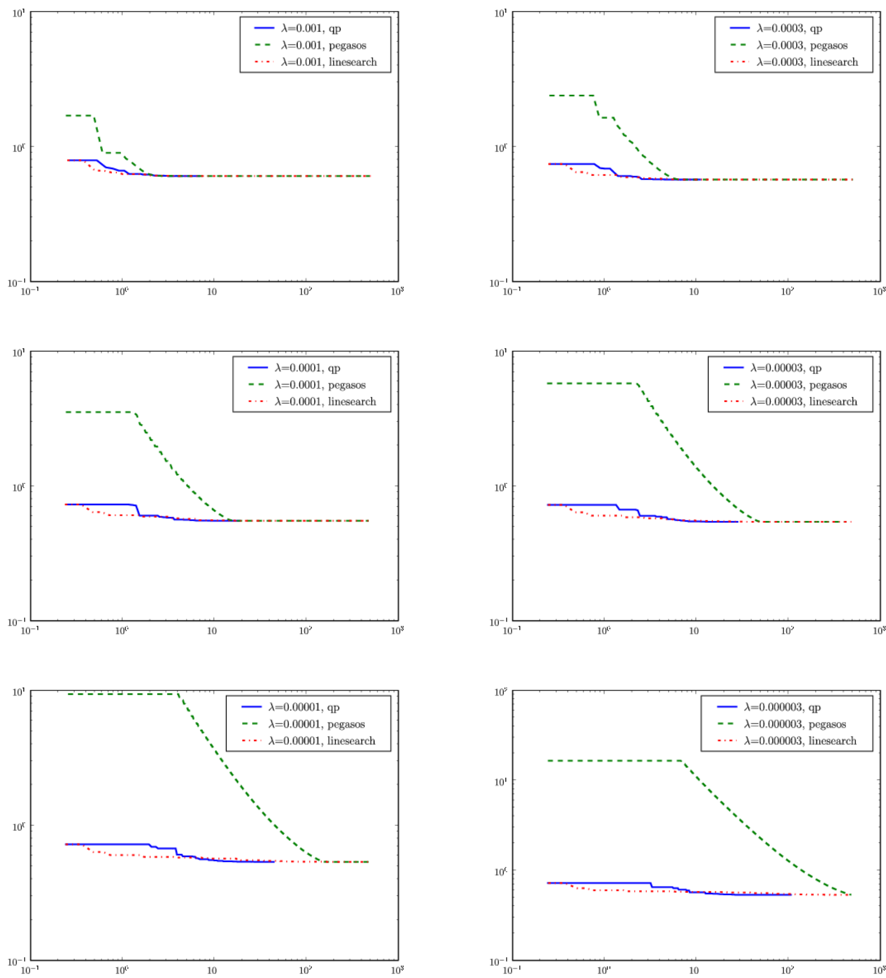

Figure 10: Objective function value with respect to time on covertype with various values of λ.

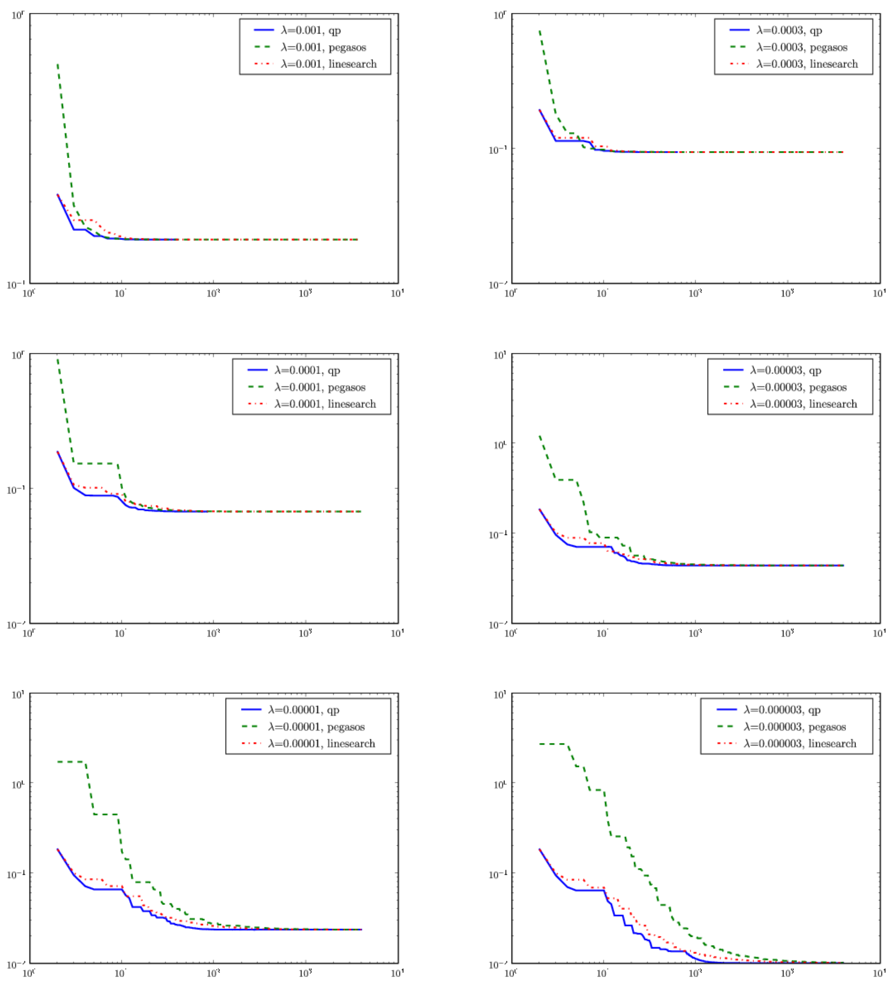

Figure 11: Objective function value with respect to number of iterations on adult9 with various values of λ.

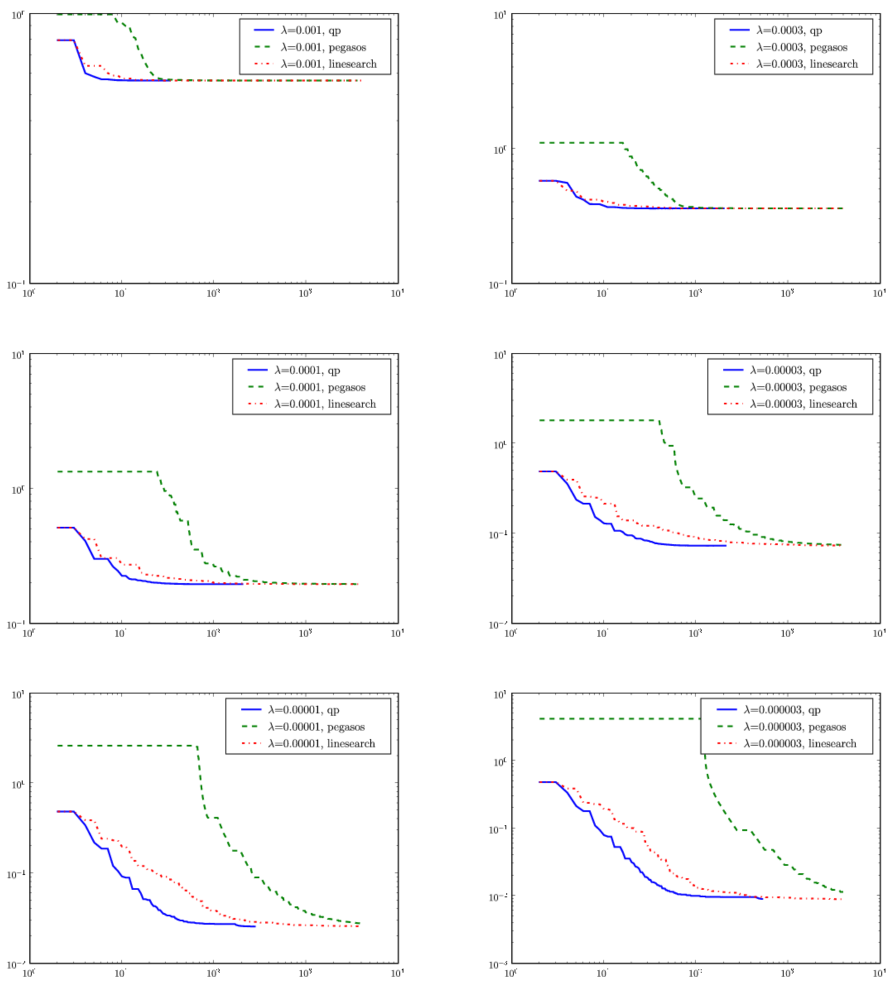

Figure 12: Objective function value with respect to number of iterations on news20 with various values of λ.

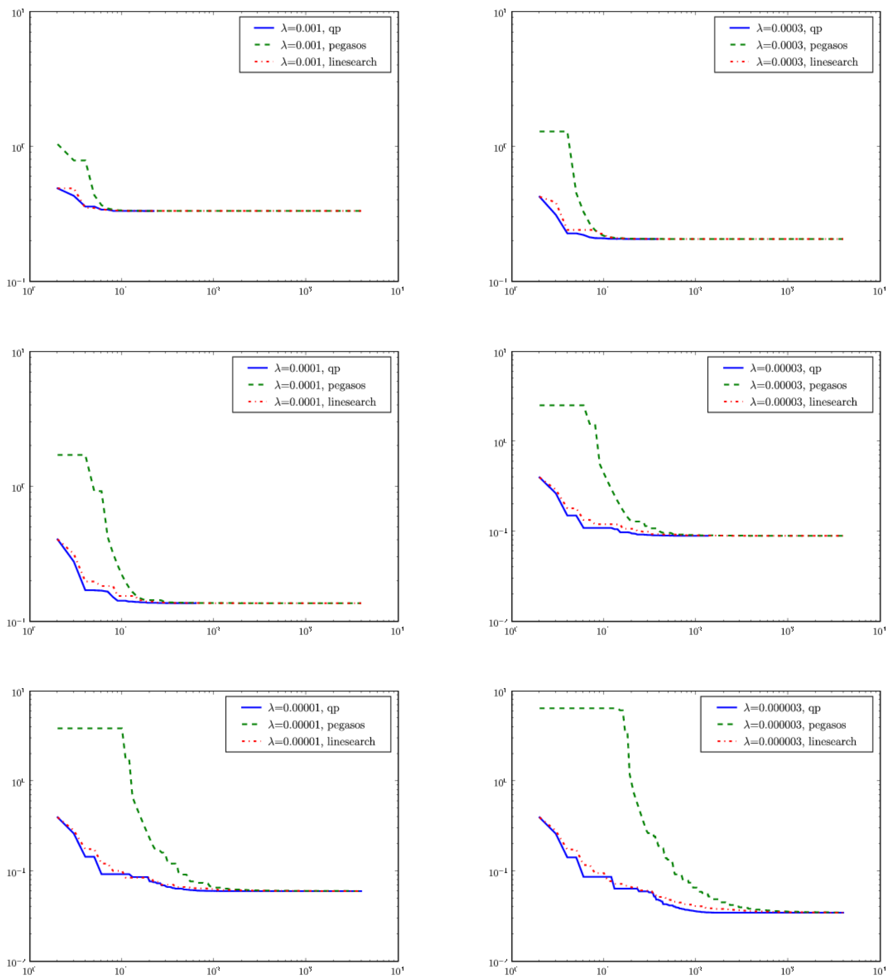

Figure 13: Objective function value with respect to number of iterations on astro-ph with various values of λ.