KAUST

Accelerated Minibatch Coordinate Descent

Filip Hanzely, Peter Richtarik

Outline

- Introduction
  - Coordinate descent, Acceleration, minibatching
- Minibatch ACD
  - Algorithm and Convergence rate
  - Sketch of Analysis
- Importance minibatch sampling
  - Bounds
  - Superiority over uniform sampling
- Experiments

Coordinate Descent

Problem

$x \in \mathbb{R}^n$ and f is smooth and strongly convex

minimize $f(x)$

M smoothness

positive definite matrix

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}(x-y)^\top M (x-y)$$

σ strong convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\sigma}{2}\|x-y\|^2$$

Matrix Smoothness

More general to Lipschitz continuity of gradients (smoothness)

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2$$

$M = LI$

ϕ_i is L_i smooth

$$f(x) = \sum_{j=1}^{m} \phi_i(A_i x)$$

f is $\sum_{j=1}^{m} L_i A_i^\top A_i$ smooth

$$M = \sum_{j=1}^{m} L_i A_i^\top A_i$$

ERM with linear predictors

Logistic regression

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} \log\left(1 + \exp\left(\mathbf{A}_{i,:}x \cdot b\right)\right) + \frac{\lambda}{2}\|x\|^2$$

Dual of SVM with squared hinge loss

$$f(x) = \frac{1}{\lambda n^2}\sum_{j=1}^{m}\left(\sum_{i=1}^{n} b_i \mathbf{A}_{ji} x_i\right)^2 - \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{4n}\sum_{i=1}^{n} x_i^2 + \mathcal{I}_{[0,\infty]}(x)$$

Indicator function - proximable

$n > m \quad \rightarrow \quad$ CD is the state-of-the-art

Randomized Coordinate Descent

Pick randomly subset of coordinates, take a gradient step on them

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}(x-y)^\top M (x-y)$$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\sigma}{2}\|x-y\|^2$$

Gradient Descent $\Rightarrow$ $\left(1 - \frac{\sigma}{\lambda_{\max}(M)}\right)^k$

Coordinate Descent (Importance Sampling) $\Rightarrow$ $\left(1 - \frac{\sigma}{\mathrm{Trace}(M)}\right)^k$

n times cheaper

$p \propto \mathrm{Diag}(M)$

Acceleration [Nesterov 1983]

"Square rooting" condition number for gradient descent

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2$$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\sigma}{2}\|x - y\|^2$$

Gradient Descent → $\left(1 - \frac{\sigma}{L}\right)^k$

Accelerated GD → $\left(1 - \sqrt{\frac{\sigma}{L}}\right)^k$

Matches lower bound

Accelerated Coordinate Descent [Allen-Zhu et al. 2016]

Coordinate Descent → $\left(1 - \frac{\sigma}{\mathrm{Trace}(M)}\right)^k$

Accelerated Coordinate Descent → $\left(1 - \frac{\sqrt{\sigma}}{\sum \sqrt{M_{i,i}}}\right)^k$

$p \propto \mathrm{Diag}(M)^{\frac{1}{2}}$

Previous ACD algorithms give a worse rate

Minibatching

Analyzed in non-accelerated case

Smoothness in expectation

Consequence of M smoothness

Expected Separable Overapproxiamtion (ESO)

$$\forall x, h: \quad \mathbb{E}\left[f\left(x+\sum_{i\in S} h_i e_i\right)\right] \leq f(x) + \sum_{i=1}^{n} p_i \nabla_i f(x) h_i + \frac{1}{2}\sum_{i=1}^{n} p_i v_i h_i^2$$

probability vector, $P(i \in S) = p_i$

"ESO" vector v

$v = \mathrm{diag}(M)$ when sampling one coordinate at time

ESO

probability matrix

$$\mathbf{P}_{i,j} = P(i \in S \wedge j \in S)$$

M smoothhess

$+$

$$\mathbf{P} \circ M \preceq \mathrm{Diag}(p_1 v_1, \ldots, p_n v_n)$$

$$\forall x, h: \quad \mathbb{E}\left[f\left(x + \sum_{i \in S} h_i e_i\right)\right] \leq f(x) + \sum_{i=1}^{n} p_i \nabla_i f(x) h_i + \frac{1}{2} \sum_{i=1}^{n} p_i v_i h_i^2$$

Contributions

- Accelerated minibatch coordinate descent with arbitrary probabilities
- Importance minibatch sampling for CD
- Importance minibatch sampling for ACD

Minibatch ACD

Algorithm

$$\sigma_w = \min_i \frac{p_i^2 \sigma}{v_i}$$

$$\theta \approx 0.618\,\sigma_w$$

$$x^{k+1} = (1-\theta)y^k + \theta z^k$$

$$y^{k+1} = x^{k+1} - \sum_{i \in S^k} \frac{1}{v_i} \nabla_i f(x^{k+1}) e_i$$

$$z^{k+1} = \frac{1}{1+\eta\sigma_w} \left(z^k + \eta\sigma_w x^{k+1} - \sum_{i \in S^k} \frac{\eta}{p_i} \nabla_i f(x^{k+1}) e_i \right)$$

$$\eta \approx 1.618\,\sigma_w^{-\frac{1}{2}}$$

Algorithm and rate

$$x^{k+1} = (1-\theta)y^k + \theta z^k$$

$$y^{k+1} = x^{k+1} - \sum_{i \in S^k} \frac{1}{v_i} \nabla_i f(x^{k+1}) e_i$$

$$z^{k+1} = \frac{1}{1+\eta\sigma_w} \left(z^k + \eta\sigma_w x^{k+1} - \sum_{i \in S^k} \frac{\eta}{p_i} \nabla_i f(x^{k+1}) e_i \right)$$

"Arbitrary sampling" result

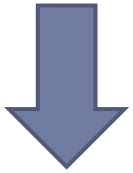
$$\Psi^k = \frac{1}{\theta^2}(f(y^k) - f(x^*)) + \frac{1}{2(1-\theta)}\|z^k - x^*\|^2$$

$$\mathbb{E}[\Psi^k] \leq \left(1 - \sqrt{\sigma \min_i \frac{p_i^2}{v_i}}\right)^k \Psi^0$$

Recovered results from Allen-Zhu et al (2016) as special case

Sketch of Analysis [Allen-Zhu et al 2014]

$$y^{k+1} = x^{k+1} - \sum_{i \in S^k} \frac{1}{v_i} \nabla_i f(x^{k+1}) e_i$$



$$f(x^{k+1}) - \mathbb{E}[f(y^{k+1}) \,|\, x^{k+1}] \geq \frac{1}{2} \|\nabla f(x^{k+1})\|^2_{v^{-1} \circ p}$$

Gradient descent lemma

Sketch of Analysis [Allen-Zhu et al 2014]

$$z^{k+1} = \frac{1}{1+\eta\sigma}\left(z^k + \eta\sigma x^{k+1} - \sum_{i \in S^k} \frac{\eta}{p_i} \nabla_i f(x^{k+1}) e_i \right)$$

$$\eta \sum_{i \in S^k} \left\langle \frac{1}{p_i} \nabla_i f(x^{k+1}) e_i, z^{k+1} - u \right\rangle - \frac{\eta\sigma}{2} \|x^{k+1} - u\|^2$$
$$\leq -\frac{1}{2}\|z^k - z^{k+1}\|^2 + \frac{1}{2}\|z^k - u\|^2 - \frac{1+\eta\sigma}{2}\|z^{k+1} - u\|^2.$$

Dual averaging lemma

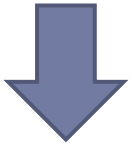
Importance minibatch sampling

Bound on convergence rate

$\mathbb{E}[|S|] = \tau$

CD: no better than $\left(1 - \tau \frac{\sigma}{\text{Trace}(M)}\right)^k$

ACD: no better than $\left(1 - \tau \frac{\sqrt{\sigma}}{\sum_{i=1}^n M_{i,i}^{\frac{1}{2}}}\right)^k$



No superlinear speedup

Importance sampling for minibatches

τ-nice sampling: $|S| = \tau$, each subset with equal probability

A little work on importance minibatch samplings

Csiba et al 2016: "Bucket sampling"

No sampling is globally better to uniform

We can almost (up to constant factor) establish it

Samplings

Independent sampling: $i \in S$ and $j \in S$ are independent

τ-nice sampling: $|S| = \tau$, each subset with equal probability

Lemma: τ-nice sampling is at most $\frac{1}{1-\frac{n-\tau}{n(n-1)}}$ times better to independent uniform sampling of expected size τ

Similar result for accelerated case

Convergence rates recap

Coordinate Descent → $\left(1 - \sigma \min_i \frac{p_i}{v_i}\right)$ $v \propto p$

Accelerated Coordinate Descent → $\left(1 - \sqrt{\sigma \min_i \frac{p_i^2}{v_i}}\right)$ $v \propto p^2$

Importance Sampling for CD

p is almost proportional to $\mathrm{diag}(M)$

$$p_{\mathrm{imp}} = \frac{\mathrm{diag}(M)}{c_\tau + \mathrm{diag}(M)}$$

c_τ s.t. $\mathbb{E}(|S|) = \tau$

Lemma: Independent uniform sampling is at most $\frac{2n-\tau}{n-\tau}$ times better

τ nice sampling is cannot be much better to importance

Importance Sampling for CD

τ nice sampling is cannot be much better to importance

Example 2. Consider $n \gg O(1)$, $\tau = 2$ and

$$M = \begin{pmatrix} n & 0^\top \\ 0 & I \end{pmatrix}$$

$I \in \mathbb{R}^{n-1,n-1}$

Importance is $\Theta(n)$ times better

Importance Sampling for ACD

almost $p \propto \sqrt{\text{diag}(M)}$

$$\frac{p}{\text{diag}(M)} \propto \frac{1}{p} - 1$$

Lemma: Independent uniform sampling is at most $\mathcal{O}(\sqrt{\tau})$ times better

τ nice sampling is cannot be much better to importance

Importance Sampling for ACD

$\mathcal{O}(\sqrt{\tau})$

τ nice sampling is cannot be much better to importance

sufficiently big

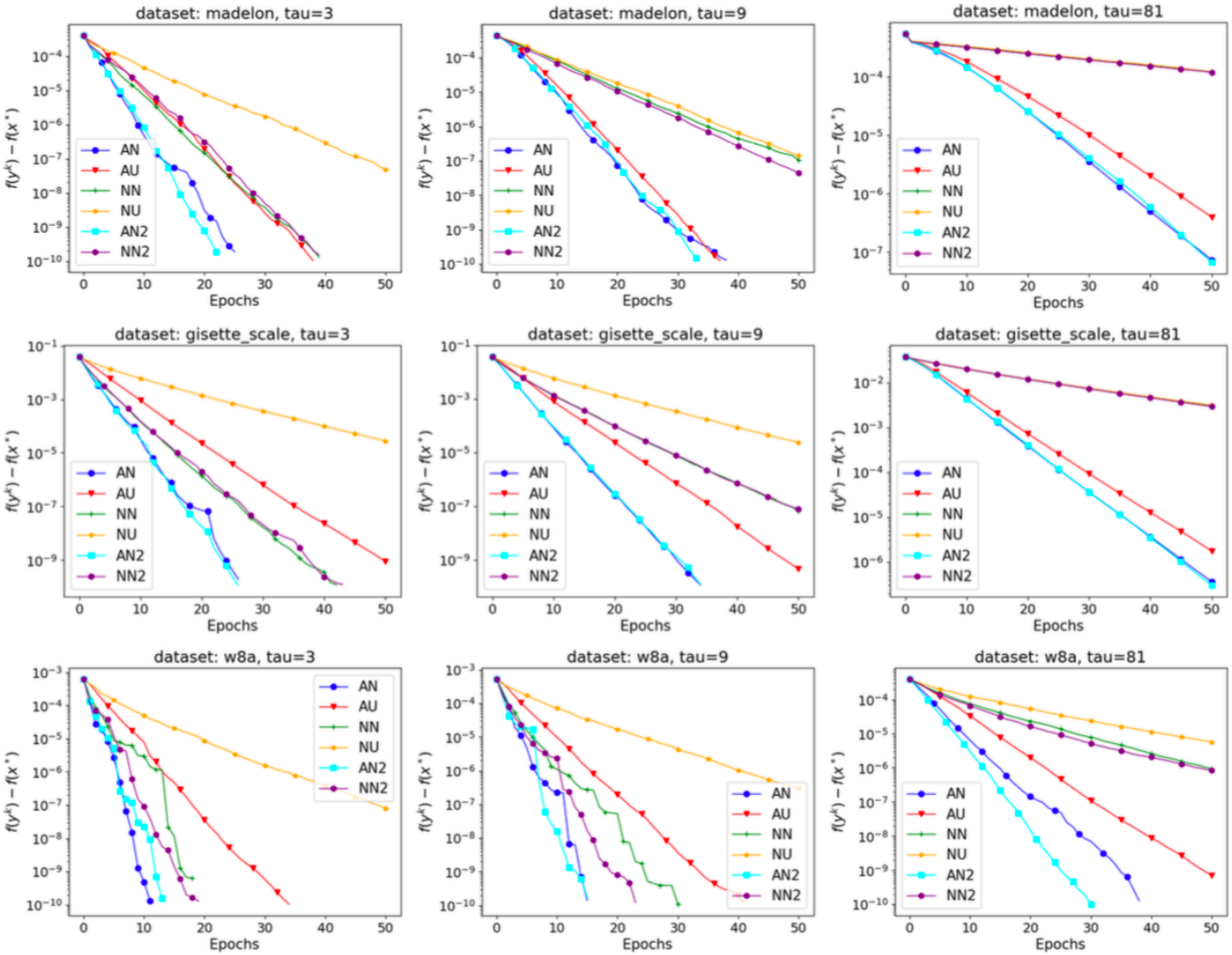
$$M = \begin{pmatrix} C & 0 \\ 0 & I \end{pmatrix}$$

$I \in \mathbb{R}^{n-1,n-1}$

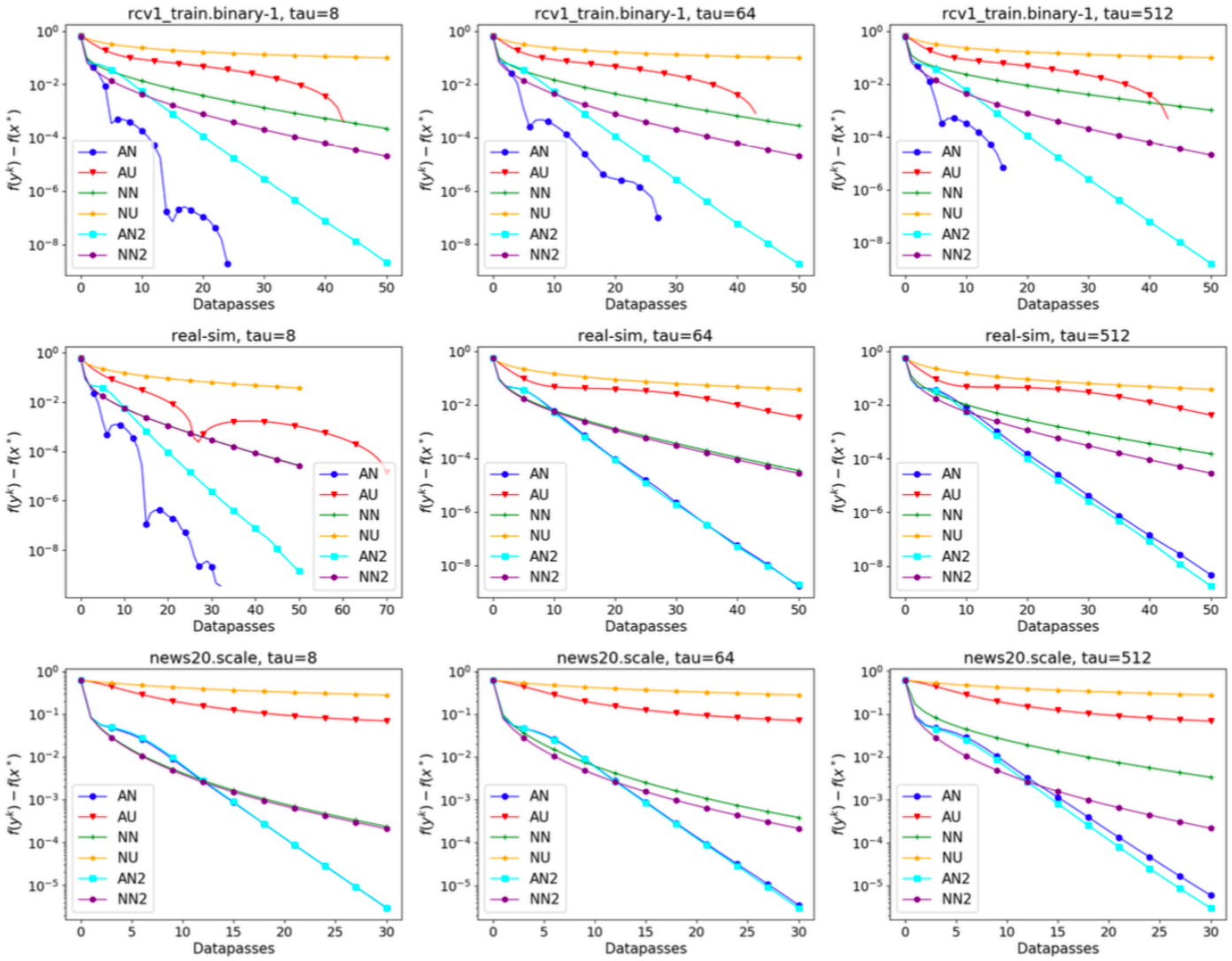
Importance is $\Theta(n)$ times better

Experiments

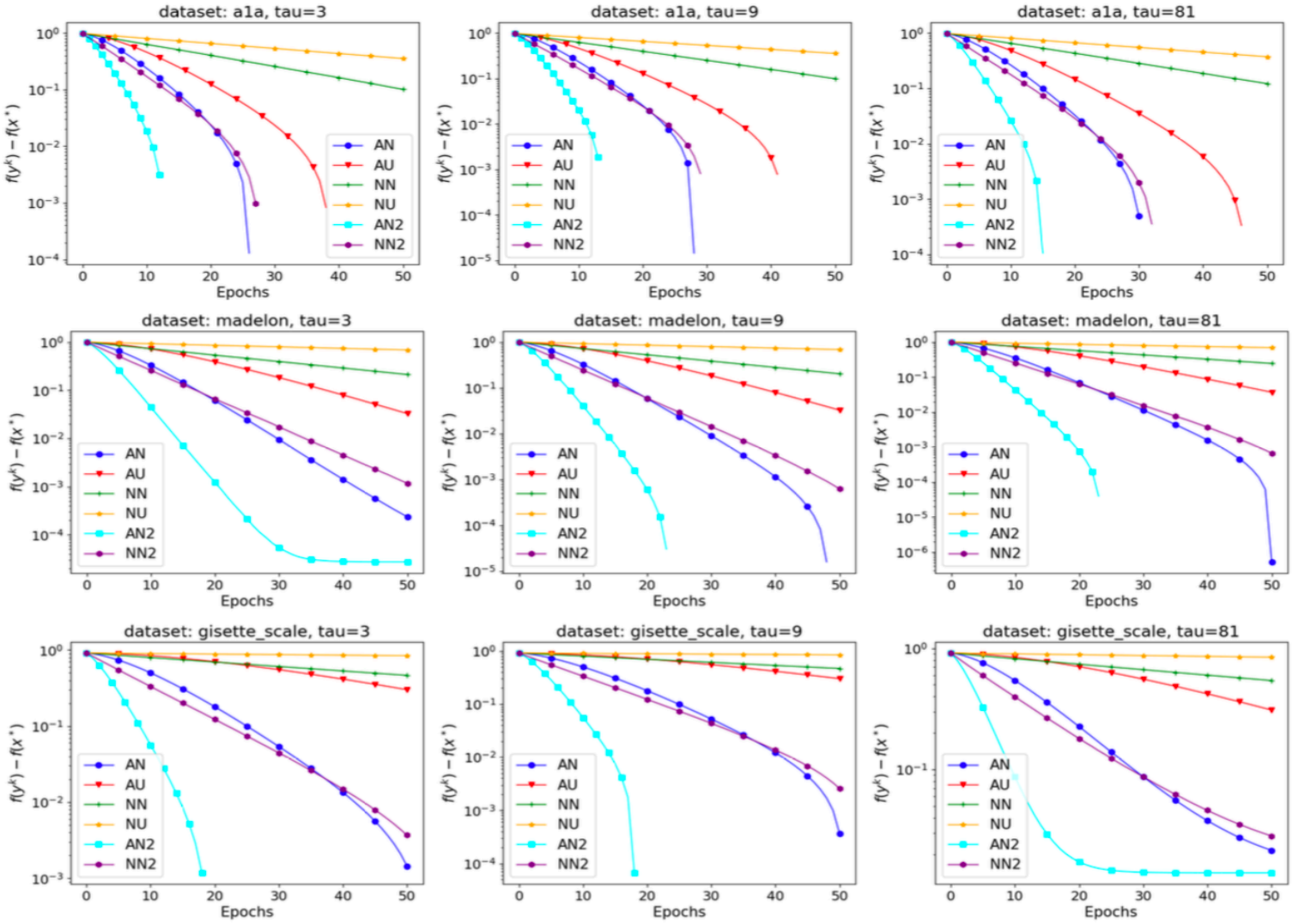
Logistic regression



Logistic regression (larger and practical)



SVM


dataset: a1a, tau=3
dataset: a1a, tau=9
dataset: a1a, tau=81
dataset: madelon, tau=3
dataset: madelon, tau=9
dataset: madelon, tau=81
dataset: gisette_scale, tau=3
dataset: gisette_scale, tau=9
dataset: gisette_scale, tau=81
$f(y^k) - f(x^*)$
Epochs
AN
AU
NN
NU
AN2
NN2