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

\[ x \in \mathbb{R}^n \text{ and } f \text{ is smooth and strongly convex} \]

\[ \text{minimize } f(x) \]

\[ M \text{ smoothness} \]

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

\[ \sigma \text{ strong convexity} \]

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

More general to Lipschitz continuity of gradients (smoothness)

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

\( \phi_i \) is \( L_i \) smooth

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

\[ f \text{ is } \sum_{j=1}^{m} L_i A_i^T A_i \text{ smooth} \]

\[ M = \sum_{j=1}^{m} L_i A_i^T A_i \]
ERM with linear predictors

Logistic regression

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} \log (1 + \exp (A_i x \cdot b)) + \frac{\lambda}{2} \|x\|^2$$

Dual of SVM with squared hinge loss

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

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

\[
\begin{align*}
    f(x) &\leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2} (x - y)^\top M (x - y) \\
    f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\sigma}{2} \| x - y \|^2
\end{align*}
\]

Gradient Descent

\[
\left(1 - \frac{\sigma}{\lambda_{\text{max}}(M)}\right)^k
\]

Coordinate Descent (Importance Sampling)

\[
\left(1 - \frac{\sigma}{\text{Trace}(M)}\right)^k
\]

\(n\) times cheaper

\(p \propto \text{Diag}(M)\)
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  \to \left(1 - \frac{\sigma}{L}\right)^k \quad \text{Matches lower bound}

Accelerated GD  \to \left(1 - \sqrt{\frac{\sigma}{L}}\right)^k
Accelerated Coordinate Descent [Allen-Zhu et al. 2016]

Coordinate Descent \[ (1 - \frac{\sigma}{\text{Trace}(M)})^k \]

Accelerated Coordinate Descent \[ (1 - \frac{\sqrt{\sigma}}{\sum \sqrt{M_{i,i}}})^k \]

\[ p \propto \text{Diag}(M)^{\frac{1}{2}} \]

Previous ACD algorithms give a worse rate
Minibatching

Analyzed in non-accelerated case

Smoothness in expectation  Consequence of $M$ smoothness

Expected Separable Overapproximation (ESO)

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

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

“ESO” vector $v$

$v = \text{diag}(M)$ when sampling one coordinate at time
proof of convergence
Contributions

- Accelerated minibatch coordinate descent with arbitrary probabilities
- Importance minibatch sampling for CD
- Importance minibatch sampling for ACD
Minibatch ACD
\[ \sigma_w = \min \frac{p_i^2 \sigma}{v_i} \]

\[ \theta \approx 0.618 \sigma_w \]

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

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

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

\[ \eta \approx 1.618 \sigma_w^{-\frac{1}{2}} \]
Algorithm and rate

\[ x^{k+1} = (1 - \theta) y^k + \theta z^k \]
\[ y^{k+1} = x^{k+1} - \sum_{i \in S_k} \frac{1}{v_i} \nabla_i f(x^{k+1}) e_i \]
\[ z^{k+1} = \frac{1}{1 + \eta \sigma_w} \left( \underbrace{z^k + \eta \sigma_w x^{k+1} - \sum_{i \in S_k} \frac{\eta}{p_i} \nabla_i f(x^{k+1}) e_i}_{\Psi^k} \right) \]

"Arbitrary sampling" result

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

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

Recovered results from Allen-Zhu et al (2016) as special case
Sketch of Analysis [Allen-Zhu et al 2014]

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

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

Gradient descent lemma
Sketch of Analysis [Allen-Zhu et al 2014]

\[
\begin{align*}
    z^{k+1} &= \frac{1}{1 + \eta\sigma} \left( z^k + \eta\sigma x^{k+1} - \sum_{i \in S_k} \frac{\eta}{p_i} \nabla_i f(x^{k+1}) e_i \right) \\
    \eta \sum_{i \in S_k} \left< \frac{1}{p_i} \nabla_i f(x^{k+1}) e_i, z^{k+1} - u \right> - \frac{\eta\sigma}{2} \|x^{k+1} - u\|^2 \\
    &\leq -\frac{1}{2} \|z^k - z^{k+1}\|^2 + \frac{1}{2} \|z^k - u\|^2 - \frac{1 + \eta\sigma}{2} \|z^{k+1} - u\|^2.
\end{align*}
\]

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

\( \tau \)-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

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

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

Similar result for accelerated case
Convergence rates recap

Coordinate Descent

\[ (1 - \sigma \min_i \frac{p_i}{v_i}) \]

\[ \nu \propto p \]

Accelerated Coordinate Descent

\[ (1 - \sqrt{\sigma \min_i \frac{p_i^2}{v_i}}) \]

\[ \nu \propto p^2 \]
Importance Sampling for CD

\[ p \text{ is almost proportional to } \text{diag}(M) \]

\[ p_{\text{imp}} = \frac{\text{diag}(M)}{c_\tau + \text{diag}(M)} \]

\[ c_\tau \text{ s.t. } \mathbb{E}(|S|) = \tau \]

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

\( \tau \) 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

\( \tau \) nice sampling is cannot be much better to importance
Importance Sampling for ACD

\[ O(\sqrt{\tau}) \]

\( \tau \) nice sampling is cannot be much better to importance

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

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

Importance is \( \Theta(n) \) times better
Experiments
Logistic regression
Logistic regression (larger and practical)
SVM