Stochastic Proximal Gradient Descent with Acceleration Techniques

Atsushi Nitanda
Presenter: Zhe Li

February 19, 2016

Main idea

- Proximal gradient descent
- Nesterov's accelation technique
- Reduced variance technique
- Mini-batch setting

Algorithm 1 Nesterov's Acceleration update

Input: total time T
Initialize: $x_1 = y_1$.
for $s = 1, 2, \cdots$ **do**
    $y_{s+1} = x_s - \frac{1}{\beta}\nabla f(x_s)$
    $x_{s+1} = (1 + \frac{\sqrt{Q}-1}{\sqrt{Q}+1})y_{s+1} - \frac{\sqrt{Q}-1}{\sqrt{Q}+1}y_s$
end for
Return: y_{t+1}

Theorem

Let f be α-strongly convex and β-smooth, then Nesterov's Accelerated Gradient Descent Satisfies

$$f(y_t) - f(x^*) \leq \frac{\alpha + \beta}{2}||x_1 - x^*||^2 \exp(-\frac{t-1}{\sqrt{Q}})$$

Proof Technique

- Define α-strongly convex quadratic function Φ_s
  - $\Phi_1(x) = f(x_1) + \frac{\alpha}{2}||x - x_1||^2$
  - $\Phi_{s+1}(x) = (1 - \frac{1}{\sqrt{Q}})\Phi_s(x) + \frac{1}{\sqrt{Q}}(f(x_s) + \nabla f(x_s)^T(x - x_s) + \frac{\alpha}{2}||x - x_s||^2)$

- $\Phi_{s+1}(x) \leq f(x) + (1 - \frac{1}{\sqrt{Q}})^s(\Phi_1(x) - f(x))$
- $f(y_s) \leq \underset{x \in \mathbb{R}^n}{min} \Phi_s(x)$

- Problem setting

$$\min P(w), P(w) := \frac{1}{n}\sum_{i=1}^{n}\psi_i(w)$$

- Smooth Assumption

$$\psi_i(w) - \psi_i(w') - \frac{L}{2}||w - w'||_2^2 \leq \nabla\psi_i(w')^T(w - w')$$

- Strongly Convex Assumption

$$P(w) - P(w') - \frac{\gamma}{2}||w - w'||_2^2 \geq \nabla P(w')^T(w - w')$$

Reduced Variance Technique

Algorithm 2 SVRG Technique

Input: update frequency m and learning rate η

Initialize: $\tilde{w}_0$.

for $s = 1, 2, \cdots$ **do**

  $\tilde{w} = \tilde{w}_{s-1}$

  $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} \nabla\psi_i(\tilde{w})$

  $w_0 = \tilde{w}$

  for $t = 1, 2 \cdots, m$ **do**

    Randomly pick $i_t \in 1, \cdots, n$ and update weight

    $w_t = w_{t-1} - \eta(\nabla\psi_i(\tilde{w}_{t-1}) - \nabla\psi_i(\tilde{w}) + \tilde{\mu})$

  end for

end for

Return: x_t

Theorem

Consider SVRG, assume that all ψ_i are convex and satisfy the two assumption with $\gamma > 0$. Let $w_* = \underset{w}{argmin} P(w)$. Assume that m is sufficiently large so that

$$\alpha = \frac{1}{\gamma\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1$$

then we have geometric convergence in expectation for SVRG

$$E[P(\tilde{w}_s)] \leq E[P(w_*)] + \alpha^s [P(\tilde{w}_0) - P(w_*)]$$

Proof Technique

- $g_i(w) = \psi_i(w) - \psi_i(w_*) - \nabla\psi_i(w_*)^T(w - w_*)$
- $n^{-1}\sum_{i=1}^{n} ||\nabla\psi_i(w) - \nabla\psi_i(w_*)||_2^2 \leq 2L[P(w) - P(w_*)]$
- $v_t = \nabla\psi_{i_t}(w_{t-1}) - \nabla\psi_{i_t}(\tilde{w}) - \tilde{\mu}$
- $E[||v_t||_2^2] \leq 4L[P(w_{t-1}) - P(w_*) + P(\tilde{w}) - P(w_*)]$
- $E[||w_t - w_*||^2] \leq ||w_{t-1} - w_*||^2 - 2\eta(1 - 2L\eta)[P(w_{t-1}) - P(w_*)] + 4L\eta^2[P(\tilde{w}) - P(w_*)]$
- $E[||w_m - w_*||^2]||^2] + 2\eta(1 - 2L\eta)mE[P(\tilde{w}_s) - P(w_*)] \leq 2(\gamma^{-1} + 2Lm\eta^2)E[P(\tilde{w}) - P(w_*)]$