Convergence diagnostics for stochastic gradient descent with constant learning rate

Supplementary material

Theorem 1 ([3, 4]) *Under certain assumptions on the loss function, there are positive constants* A_γ, B *such that, for every* n*, it holds that*

$$\mathbb{E}(||\theta_n - \theta_\star||^2) \leq \mathbb{E}(||\theta_0 - \theta_\star||^2)e^{-A_\gamma n} + B\gamma.$$

Theorem 2 *Consider SGD with constant rate,*

$$\theta_n = \theta_{n-1} - \gamma \nabla \ell(y_n, x_n^\top \theta_{n-1}).$$

Suppose that Theorem 1 holds, so that that $\mathbb{E}(||\theta_n - \theta_\star||^2) \leq \gamma M$*, for some positive* M *and large enough* n*. We make the following additional assumptions:*

(a) $\nabla \ell(y, x^\top \theta) = f(x, \theta) + e$*, where* $f(x, \theta)$ *is L-Lipschitz,* $\mathbb{E}(e|x, \theta) = 0$ *and* $\mathbb{E}(||e||^2) \geq \tau^2$*.*

(b) It holds $\mathbb{E}(f(x, \theta - \gamma z)^\top z) \leq \mathbb{E}(f(x, \theta)^\top z) - \gamma K \cdot \mathbb{E}(z^\top C z)$*, for any* θ, z*, for some positive constant* K*, and some positive definite matrix* C *with minimum eigenvalue* $\mu > 0$*.*

(c) It holds that $\gamma > (L^2 M - \mu K \tau^2)/\mu K L^2 M$*.*

Then,

$$\mathbb{E}(\nabla \ell(y_n, x_n^\top \theta_{n-1})^\top \nabla \ell(y_{n+1}, x_{n+1}^\top \theta_n)) < 0.$$

Proof 2 *For brevity let* $\tilde{\ell}_i = f(x_{i+1}, \theta_i) + e_i = f_i + e_i$ *be the stochastic gradient at iteration* $i + 1$*.*

$$\begin{aligned}
\mathbb{E}(\tilde{\ell}_{i-1}^\top \tilde{\ell}_i) &= \mathbb{E}\left[(f_{i-1} + e_{i-1})^\top (f_i + e_i)\right] = \mathbb{E}\left[(f_{i-1} + e_{i-1})^\top f_i\right] \qquad \textit{[because } e_i \textit{ are zero-mean]} \\
&= \mathbb{E}\left[(f_{i-1} + e_{i-1})^\top f\left(\theta_{i-1} - \gamma f_{i-1} - \gamma e_{i-1}\right)\right] \qquad \textit{[by SGD step for } \theta_i \textit{]} \\
&\leq \mathbb{E}(||f_{i-1}||^2) - \gamma K \cdot \mathbb{E}\left[(f_{i-1} + e_{i-1})^\top C (f_{i-1} + e_{i-1})\right] \qquad \textit{[by Assumption (b)]} \\
&\leq (1 - \gamma \mu K)\mathbb{E}(||f_{i-1})||^2) - \gamma K \cdot \mathbb{E}(||e_{i-1}||_C^2) \\
&\leq (1 - \gamma \mu K) L^2 \mathbb{E}(||\theta_{i-1} - \theta_\star||^2) - \gamma \mu K \tau^2 \qquad \textit{[by Lipschitz Assumption (a)]} \\
&\leq \gamma[(1 - \gamma \mu K) L^2 M - \mu K \tau^2] \\
&< 0. \qquad \textit{[by Assumption (c) and small enough } \gamma \textit{]}
\end{aligned} \tag{1}$$

Remarks. Assumption (b) is a form of strong convexity. For example, suppose that $y = x^\top \theta_\star + e$, then $f(x, \theta) = x x^\top (\theta - \theta_\star)$ and $f(x, \theta - \gamma z)^\top z = f(x, \theta)^\top z - \gamma z^\top \mathbb{E}(x x^\top) z$. In this case $C = \mathbb{E}(x x^\top)$

is the Fisher information matrix and Assumption (b) holds for $K = 1$. When γ is small enough and a Taylor approximation of $f(x, \theta - \gamma z)$ is possible, the above result still holds for $K = 1$ when the Fisher information exists. Assumption (c) shows that there is a threshold value for γ below which the diagnostic cannot terminate. For example, suppose that error noise is small so that $\tau^2 \approx 0$ and $K = 1$, as argued before. Then, $\gamma > 1/\mu$, that is, the learning rate has to exceed the reciprocal of the minimum eigenvalue of the Fisher information matrix.

Theorem 3 *Suppose that the loss is quadratic, $\ell(y, x^\top\theta) = (1/2)(y - x^\top\theta)^2$. Let x_1 and x_2 be two iid vectors from the distribution of x, and define: $\sigma^2 = \mathbb{E}((y - x^\top\theta_\star)^2)$; $c^2 = \mathbb{E}((x_1^\top x_2)^2)$; $C = \mathbb{E}(x_1 x_2^\top(x_1^\top x_2))$; $D = \mathbb{E}(x_1 x_1^\top(x_1^\top x_2)^2)$, and suppose that all such constants are finite. Then, for $\gamma > 0$,*

$$\begin{aligned}\Delta_n(\theta) &= \mathbb{E}(S_{n+2} - S_{n+1} | \theta_n = \theta) \\ &= (\theta - \theta_\star)^\top(C - \gamma D)(\theta - \theta_\star) - \gamma c^2\sigma^2.\end{aligned}$$

Proof 3 *For notational brevity we make the following definitions:*

$$\begin{aligned}\theta^+ &= \theta + \gamma(y_1 - x_1^\top\theta)x_1 \\ \theta^{++} &= \theta^+ + \gamma(y_2 - x_2^\top\theta^+)x_2,\end{aligned} \tag{2}$$

where θ is the current iterate, and θ^+ and θ^{++} are the next two using iid data (x_1, y_1) and (x_2, y_2). For a fixed θ we understand the Pflug diagnostic through the function

$$H(\theta) = S_{++} - S_+|\theta = \nabla_{++}\ell^\top\nabla_+\ell = (\theta^+ - \theta)^\top(\theta^{++} - \theta^+)/\gamma^2 \tag{3}$$

$$\text{and } \Delta_n(\theta) = \mathbb{E}(H(\theta)) = \mathbb{E}\left((\theta^+ - \theta)^\top(\theta^{++} - \theta^+)/\gamma^2\right). \tag{4}$$

We use Equation (2) to derive an expression for H:

$$\begin{aligned}H(\theta) &= (y_1 - x_1^\top\theta)(y_2 - x_2^\top\theta^+)x_1^\top x_2 \\ &= (y_1 - x_1^\top\theta)\left[y_2 - x_2^\top\theta - \gamma(y_1 - x_1^\top\theta)x_1^\top x_2\right]x_1^\top x_2 \\ &= (y_1 - x_1^\top\theta)(y_2 - x_2^\top\theta)x_1^\top x_2 - \gamma(y_1 - x_1^\top\theta)^2(x_1^\top x_2)^2.\end{aligned} \tag{5}$$

Let $y_i = x_i^\top\theta_\star + \varepsilon_i$; we know that $\mathbb{E}((y_i - x_i^\top\theta_\star)x_i) = 0$. Now, we analyze each term individually:

$$\begin{aligned}(y_1 - x_1^\top\theta)(y_2 - x_2^\top\theta)x_1^\top x_2 &= [x_1^\top(\theta_\star - \theta) + \varepsilon_1][x_2^\top(\theta_\star - \theta) + \varepsilon_2]x_1^\top x_2 \\ &= (\theta - \theta_\star)^\intercal x_1 x_2^\top(x_1^\top x_2)(\theta - \theta_\star) + \varepsilon_1 W^{(1)} + \varepsilon_2 W^{(2)} + \varepsilon_1\varepsilon_2 W^{(3)}.\end{aligned}$$

The W variables are conditionally independent of ε and so using the law of iterated expectations these terms vanish.

$$\mathbb{E}\left((y_1 - x_1^\top\theta)(y_2 - x_2^\top\theta)x_1^\top x_2\right) = (\theta - \theta_\star)^\intercal\mathbb{E}\left(x_1 x_2^\top(x_1^\top x_2)\right)(\theta - \theta_\star) = (\theta - \theta_\star)^\intercal C(\theta - \theta_\star).$$

Using a similar reasoning, for the second term we have:

$$\begin{aligned}(y_1 - x_1^\top\theta)^2(x_1^\top x_2)^2 &= \left[(x_1^\top(\theta_\star - \theta) + \varepsilon_1\right]^2(x_1^\top x_2)^2 \\ &= (\theta - \theta_\star)^\intercal x_1 x_1^\top(x_1^\top x_2)^2(\theta - \theta_\star) + \varepsilon_1 W^{(4)} + \varepsilon_1^2(x_1^\top x_2)^2.\end{aligned} \tag{6}$$

In expectation of Equation (6),

$$\mathbb{E}\left((y_1 - x_1^\top\theta)^2(x_1^\top x_2)^2\right) = (\theta-\theta_\star)^\intercal \mathbb{E}(x_1 x_1^\top (x_1^\top x_2)^2)(\theta-\theta_\star) + \varepsilon_1^2(x_1^\top x_2)^2$$
$$= (\theta-\theta_\star)^\intercal D(\theta-\theta_\star) + \sigma^2 c^2. \tag{7}$$

By combining all results we finally get:

$$\Delta_n(\theta) = (\theta-\theta_\star)^\intercal (C-\gamma D)(\theta-\theta_\star) - \gamma\sigma^2 c^2.$$

Theorem 4 *Let* $\lambda_\gamma = \mathbb{E}(1/(1+\gamma||x||^2)) \in (0,1]$. *Under the assumptions of Theorem 3 applied on the implicit procedure in Equation* (9)*, it holds that*

$$\Delta_n^{\text{im}}(\theta) = \mathbb{E}(S_{n+2} - S_{n+1}|\theta_n = \theta)$$
$$= a_\gamma \Delta_n(\theta) + b_\gamma \left[(\theta-\theta_\star)^\top D(\theta-\theta_\star) + \sigma^2 c^2\right],$$

where $a_\gamma = \lambda_\gamma^2$, $b_\gamma = \gamma\lambda_\gamma^2(1-\lambda_\gamma)$.

Proof 4 *We derive similar theoretical results for* $H^{im}(\theta), \Delta_n^{im}(\theta)$ *under the linear normal model for implicit updates. We have the implicit updates*

$$\theta^+ = \theta + \gamma(y_1 - x_1^\top\theta^+)x_1$$
$$\theta^{++} = \theta^+ + \gamma(y_2 - x_2^\top\theta^{++})x_2$$

Also note the collinearity

$$(y_1 - x_1^\top\theta^+) = \lambda_1(y_1 - x_1^\top\theta)$$
$$(y_2 - x_2^\top\theta^{++}) = \lambda_2(y_2 - x_2^\top\theta^+),$$
$$= \lambda_2[y_2 - x_2^\top\theta - \gamma\lambda_1(y_1 - x_1^\top\theta)x_1^\top x_2],$$

where $\lambda_1 = 1/(1+\gamma||x_1||^2)$ *and* $\lambda_2 = 1/(1+\gamma||x_2||^2)$. *We derive an expression for* H^{im}*, with implicit updates:*

$$H^{im}(\theta) = (\theta^+ - \theta)^\top(\theta^{++} - \theta^+)/\gamma^2$$
$$= (y_1 - x_1^\top\theta^+)(y_2 - x_2^\top\theta^{++})x_1^\top x_2$$
$$= \lambda_1\lambda_2(y_1 - x_1^\top\theta)[y_2 - x_2^\top\theta - \gamma\lambda_1(y_1 - x_1^\top\theta)x_1^\top x_2]x_1^\top x_2$$
$$= \lambda_1\lambda_2\left[H(\theta) + \gamma(1-\lambda_1)(y_1 - x_1^\top\theta)^2(x_1^\top x_2)^2\right]$$
$$= \lambda_1\lambda_2 H(\theta) + \gamma\lambda_1\lambda_2(1-\lambda_1)(y_1 - x_1^\top\theta)(x_1^\top x_2)^2,$$

where H *is the function from the explicit update in Equation* (5)*. The formula for* $\Delta_n^{im}(\theta)$ *follows by applying expectation and the reasoning in Equation* (7)*. Note that* $\mathbb{E}(\lambda_1\lambda_2) = \lambda_\gamma^2$ *since* λ_1 *and* λ_2 *are independent and have marginally identical distributions.*

Theorem 5 *Consider the GLM loss defined as* $\ell(y, x^\top\theta) = -y \cdot x^\top\theta + f(x^\top\theta)$. *Let* $h(u) = f'(u)$ *and suppose that* $h'(x^\top\theta) \geq k > 0$*, almost surely for all* θ. *Let* x_1, x_2 *be two iid vectors from the distribution of* x. *Define* $\sigma^2 = \mathbb{E}((y - h(x^\top\theta_\star)^2)$; $c^2 = \mathbb{E}((x_1^\top x_2)^2)$; $C(\theta,\theta_\star) = \mathbb{E}([h(x_1^\top\theta) - h(x_1^\top\theta_\star)]x_1)$; $D^2(\theta,\theta_\star) = \mathbb{E}([h(x_1^\top\theta) - h(x_1^\top\theta_\star)]^2(x_1^\top x_2)^2)$. *Then, for small enough* γ,

$$\Delta_n^{glm}(\theta) = \mathbb{E}(S_{n+2} - S_{n+1}|\theta_n = \theta)$$
$$\leq ||C(\theta,\theta_\star)||^2 - \gamma k[\sigma^2 c^2 + D^2(\theta,\theta_\star)].$$

Proof 5 *The updates for the GLM loss are as follows:*

$$\begin{aligned}\theta^{+} &= \theta + \gamma(y_1 - h(x_1^\top \theta))x_1 \\ \theta^{++} &= \theta^{+} + \gamma(y_2 - h(x_2^\top \theta^{+}))x_2,\end{aligned} \tag{8}$$

Note that $h(x_2^\top \theta^{+}) = h(x_2^\top \theta) + \gamma h'(x_2^\top \theta)(y_1 - h(x_1^\top \theta))x_2^\top x_1 + O(\gamma^2)$*. We can now follow the exact same reasoning as in Theorem 3 and that* $h'(x^\top \theta) \geq k$ *almost surely.*

1 Mean squared error bound for constant learning rate ISGD

In this section, ℓ will denote likelihood, which is the negated loss (cf. Equation (9)). Thus, we have the implicit update of SGD (ISGD):

$$\theta_n = \theta_{n-1} + \gamma \nabla \ell(y_n, x_n^\top \theta_n). \tag{9}$$

We will operate under the following assumptions:

Assumption 1 *The following assumptions are true with regard to procedure in Equation* (9)*.*

(a) Function ℓ *is convex, twice differentiable almost surely with respect to* $x^\top \theta$*.*

(b) For the observed Fisher information matrix $\hat{\mathcal{I}}_n(\theta) = \nabla^2 \ell(y_n, x_n^\top \theta)$ *there exists constants* $b > 0$ *and* $0 < t < \infty$ *such that* $b \leq trace(\hat{\mathcal{I}}_n(\theta)) \leq t$ *almost surely, for all* θ*. The Fisher information matrix* $\mathcal{I}(\theta_*) = \mathbb{E}\left(\hat{\mathcal{I}}_n(\theta_*)\right)$ *has minimum eigenvalue* $\lambda > 0$*.*

(c) There exists $\sigma^2 > 0$ *such that, for all* n*,* $\mathbb{E}(\|\nabla \ell(y_n, x_n^\top \theta_\star\|^2 | \mathcal{F}_{n-1}) \leq \sigma^2$*, almost surely.*

(d) The function $\theta \mapsto \mathbb{E}(\nabla \ell(y, x^\top \theta))$ *is Lipschitz with constant* L*, i.e., for all* n, θ_1, θ_2*,*

$$\mathbb{E}(\|\nabla \ell(y_n; x_n^\top \theta_1) - \nabla \ell(y_n; x_n^\top \theta_2)\|^2 | \mathcal{F}_{n-1}) \leq L^2 \|\theta_1 - \theta_2\|^2.$$

(e) Learning rate $\gamma > 0$ *is such that* $\gamma L^2(1 + \gamma t) < \lambda(1 + \gamma b)^2$*.*

To prove Theorem 8, our result for the upper bound on the MSE for constant learning rate ISGD, we first prove the following results:

Lemma 6 *The gradient* $\nabla \ell$ *is a scaled version of covariate* x*, i.e., for every* $\theta \in \mathbb{R}^p$ *there is a scalar* $\lambda \in \mathbb{R}$ *such that*

$$\nabla \ell(y; x^\top \theta) = \lambda x$$

Thus, the gradient in the implicit update is a scaled version of the gradient calculated at the previous iterate, i.e.,

$$\nabla \ell(y_n; x_n^\top \theta_n) = \lambda_n \nabla \ell(y_n; x_n^\top \theta_{n-1}), \tag{10}$$

where the scalar λ_n *satisfies*

$$\lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) = \ell'(y_n; x_n^\top \theta_{n-1} + \gamma \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) x_n^\top x_n) \tag{11}$$

Proof 6 *From the chain rule* $\nabla \ell(y_n; x_n^\top \theta_n) = \ell'(y_n; x_n^\top \theta_n) x_n$ *, and similarly* $\nabla \ell(y_n; x_n^\top \theta_{n-1}) = \ell'(y_n; x_n^\top \theta_{n-1}) x_n$*. Thus the two gradients are collinear. Therefore there exists a scalar* λ_n *such that*

$$\ell'(y_n; x_n^\top \theta_n) x_n = \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) x_n \tag{12}$$

We also have,

$$\begin{aligned} \theta_n &= \theta_{n-1} + \gamma \nabla \ell(y_n; x_n^\top \theta_n) \text{ [by definition of implicit SGD update Equation (9)]} \\ &= \theta_{n-1} + \gamma \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) x_n \text{ [by chain rule and Equation(12)]} \end{aligned} \tag{13}$$

Substituting the expression for θ_n *in Equation*(13) *into Equation*(12) *we obtain the desired result of the theorem. From Equation*(12) *we get the equality*

$$\ell'(y_n; x_n^\top \theta_n) = \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) \tag{14}$$

and substituting we get our desired result

$$\begin{aligned} \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) &= \ell'(y_n; x_n^\top (\theta_{n-1} + \gamma \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) x_n)) \\ &= \ell'(y_n; x_n^\top \theta_{n-1} + \gamma \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) x_n^\top x_n) \end{aligned}$$

Lemma 7 *Suppose Assumptions 1 (a), and (b) hold. Then, almost surely it holds*

$$\frac{1}{1+\gamma t} \leq \lambda_n \leq \frac{1}{1+\gamma b} \tag{15}$$

Proof 7 *From Lemma 6 we have*

$$\ell'(y_n; x_n^\top \theta_n) = \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}), \tag{16}$$

where the derivative of ℓ *is with respect to the natural parameter* $x^\top \theta$*. Using the definition of the implicit update Equation* (9),

$$\theta_n = \theta_{n-1} + \gamma \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) x_n. \tag{17}$$

We substitute this definition of θ_n *into Equation*(16) *and perform a Taylor approximation on* ℓ'*. Recall Taylor approximation for a function* f*,* $f(x) = f(a) + f'(\xi)(x-a)$ *where* ξ *lies in the closed interval between* a *and* x*. From Equation*(17) *we let* $\theta_{n-1} = a$ *and* $\gamma \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) x_n = (x-a)$*. Also, by the Chain rule* $\frac{\delta}{\delta \theta} \ell'(y; x^\top \theta) = \ell''(y; x^\top \theta) x^\top$*. Thus we obtain,*

$$\begin{aligned} \ell'(y_n; x_n^\top \theta_n) &= \ell'(y_n; x_n^\top \theta_{n-1}) + \ell''(y_n; x_n^\top \tilde{\theta}) x_n^\top \cdot \gamma \lambda_n \ell'(y_n; x_n^\top \theta_{n-1}) x_n \\ &= \ell'(y_n; x_n^\top \theta_{n-1}) + \gamma \lambda_n \ell''(y_n; x_n^\top \tilde{\theta}) \ell'(y_n; x_n^\top \theta_{n-1}) x_n^\top x_n \end{aligned} \tag{18}$$

where $\tilde{\theta} = \delta \theta_{n-1} + (1-\delta)\theta_n$ *and* $\delta \in [0,1]$*.*

By combining Equation(16) *with Equation*(18) *and cancelling out the first derivative term we get*

$$\lambda_n = 1 + \gamma \lambda_n \ell''(y_n; x_n^\top \tilde{\theta}) x_n^\top x_n$$

$$\lambda_n (1 - \gamma \ell''(y_n; x_n^\top \tilde{\theta}) \|x\|^2) = 1$$

$$\left(1 + \gamma \text{ trace}(\hat{\mathcal{I}}_n(\tilde{\theta}))\right) \lambda_n \leq 1 \text{ [where } \hat{\mathcal{I}} \text{ is the observed Fisher information]} \tag{19}$$

$$(1+\gamma b)\lambda_n \leq 1 \text{ [By Assumption 1 (b)]} \tag{20}$$

Now we get the other bound,

$$(1+\gamma t)\lambda_n \geq 1 \text{ [By Assumption 1 (b)]}$$

Theorem 8 *Suppose that Assumptions 1(a) - (e) hold. Then,*

$$\mathbb{E}(||\theta_n - \theta_\star||^2) \leq \left(1 - \frac{2\gamma\lambda}{1+\gamma t} + \frac{2\gamma^2 L^2}{(1+\gamma b)^2}\right)^n \mathbb{E}(||\theta_{n-1} - \theta_\star||^2) \tag{21}$$

$$+ \frac{\gamma\sigma^2(1+\gamma t)}{\lambda(1+\gamma b)^2 - \gamma L^2(1+\gamma t)} \tag{22}$$

Proof 8 *Starting from the implicit update (9), we have*

$$\begin{aligned}
\theta_n - \theta_* &= \theta_{n-1} - \theta_* + \gamma \nabla \ell(y_n; x_n^\top \theta_n) \\
\theta_n - \theta_* &= \theta_{n-1} - \theta_* + \gamma\lambda_n \nabla \ell(y_n; x_n^\top \theta_{n-1}) \text{ [By Lemma 6]} \\
\|\theta_n - \theta_*\|^2 &= \|\theta_{n-1} - \theta_*\|^2 \\
&\quad + 2\gamma\lambda_n(\theta_{n-1} - \theta_*)^\top \nabla \ell(y_n; x_n^\top \theta_{n-1}) \\
&\quad + \|\gamma\lambda_n \nabla \ell(y_n; x_n^\top \theta_{n-1})\|^2
\end{aligned} \tag{23}$$

To bound the last term,

$$\begin{aligned}
&\|\gamma\lambda_n \nabla \ell(y_n; x_n^\top \theta_{n-1})\|^2 \\
&\quad = \gamma^2\lambda_n^2 \|\nabla \ell(y_n; x_n^\top \theta_{n-1})\|^2 \\
&\quad = \gamma^2\lambda_n^2 \|\nabla \ell(y_n; x_n^\top \theta_{n-1}) - \nabla \ell(y_n; x_n^\top \theta_*) + \nabla \ell(y_n; x_n^\top \theta_*)\|^2 \\
&\quad \leq 2\gamma^2\lambda_n^2 \|\nabla \ell(y_n; x_n^\top \theta_{n-1}) - \nabla \ell(y_n; x_n^\top \theta_*)\|^2 + 2\gamma^2\lambda_n^2 \|\nabla \ell(y_n; x_n^\top \theta_*)\|^2 \\
&\quad \leq 2\left(\frac{\gamma}{1+\gamma b}\right)^2 \left(\|\nabla \ell(y_n; x_n^\top \theta_{n-1}) - \nabla \ell(y_n; x_n^\top \theta_*)\|^2 + \|\nabla \ell(y_n; x_n^\top \theta_*)\|^2\right) \\
&\qquad \text{[By Lemma 7]}
\end{aligned} \tag{24}$$

Taking expectation of both sides of Equation(24),

$$\begin{aligned}
&\mathbb{E}(\|\gamma\lambda_n \nabla \ell(y_n; x_n^\top \theta_{n-1})\|^2) \\
&\quad \leq 2\left(\frac{\gamma}{1+\gamma b}\right)^2 \left[\mathbb{E}(\|\nabla \ell(y_n; x_n^\top \theta_{n-1}) - \nabla \ell(y_n; x_n^\top \theta_*)\|^2) + \mathbb{E}(\|\nabla \ell(y_n; x_n^\top \theta_*)\|^2)\right] \\
&\quad \leq 2\left(\frac{\gamma}{1+\gamma b}\right)^2 \left(L^2\|\theta_{n-1} - \theta_*\|^2 + \sigma^2\right) \text{ [By Lipschitz and gradient bound, Assumption 1 (c), (d)]}
\end{aligned} \tag{25}$$

We can bound the expectation of the second term as

$$\begin{aligned}
&\mathbb{E}(2\lambda_n\gamma(\theta_{n-1} - \theta_*)^\top \nabla \ell(y_n; x_n^\top \theta_{n-1})) \\
&\quad \geq \frac{2\gamma}{1+\gamma t} \mathbb{E}\left((\theta_{n-1} - \theta_*)^\top \nabla \ell(y_n; x_n^\top \theta_{n-1})\right) \text{ [By Lemma 7]} \\
&\quad \geq \frac{2\gamma}{1+\gamma t} \mathbb{E}\left((\theta_{n-1} - \theta_*)^\top \nabla h(\theta_{n-1})\right) \text{ [where } \nabla h(\theta_{n-1}) = \mathbb{E}(\nabla \ell(y_n; x_n^\top \theta_{n-1})|\mathcal{F}_{n-1})\text{]} \\
&\quad \leq -\frac{2\gamma\lambda}{1+\gamma t} \mathbb{E}(\|\theta_{n-1} - \theta_*\|^2) \text{ [By strong convexity, Assumption 1 (b)]}
\end{aligned} \tag{26}$$

Taking expectations in (23) *and substituting inequalities* (25) *and* (26) *into* (23), *and again taking expectation, yields the recursion,*

$$\mathbb{E}(\|\theta_n - \theta_*\|^2) \leq \left(1 - \frac{2\gamma\lambda}{1+\gamma t} + \frac{2\gamma^2 L^2}{(1+\gamma b)^2}\right) \mathbb{E}(\|\theta_{n-1} - \theta_*\|^2) + 2\left(\frac{\gamma\sigma}{1+\gamma b}\right)^2 \tag{27}$$

Let $\delta_n \equiv \mathbb{E}(\|\theta_n - \theta_*\|^2)$. We can now derive the bound of the theorem as follows:

$$\delta_n \leq \left(1 - \frac{2\gamma\lambda}{1+\gamma t} + \frac{2\gamma^2 L^2}{(1+\gamma b)^2}\right)^n \delta_0 + \sum_{k=1}^{\infty} 2\left(\frac{\gamma\sigma}{1+\gamma b}\right)^2 \cdot \left(1 - \frac{2\gamma\lambda}{1+\gamma t} + \frac{2\gamma^2 L^2}{(1+\gamma b)^2}\right)^k$$
$$= \left(1 - \frac{2\gamma\lambda}{1+\gamma t} + \frac{2\gamma^2 L^2}{(1+\gamma b)^2}\right)^n \delta_0 + 2\left(\frac{\gamma\sigma}{1+\gamma b}\right)^2 \cdot \left(\frac{2\gamma\lambda}{1+\gamma t} - \frac{2\gamma^2 L^2}{(1+\gamma b)^2}\right)^{-1}$$
$$= \left(1 - \frac{2\gamma\lambda}{1+\gamma t} + \frac{2\gamma^2 L^2}{(1+\gamma b)^2}\right)^n \delta_0 + \frac{\gamma\sigma^2(1+\gamma t)}{\lambda(1+\gamma b)^2 - \gamma L^2(1+\gamma t)}$$

Lemma 9 *Suppose that Assumption 1(e) holds. The discount factor of the non-asymptotic bound in Theorem 8 will be bounded $0 < \cdot < 1$ for all $\gamma > 0$, and thus the mean squared error $\mathbb{E}(\|\theta_n - \theta_*\|^2)$ will contract for all possible values of γ. In addition the stationary term will be > 0 for all $\gamma > 0$.*

Proof 9 *The discount factor is bounded below by $\left(1 - \frac{2\gamma\lambda}{1+\gamma b} + \frac{2\gamma^2 L^2}{(1+\gamma b)^2}\right)$ because $b \leq t$. We will show that this term is bounded below by 0.*

A quick manipulation of the algebra gives us

$$\textit{(lower bound)} \quad 2\gamma\lambda(1+\gamma b) - 2\gamma^2 L^2 < (1+\gamma b)^2 \tag{28}$$
$$\textit{(upper bound)} \quad \gamma L^2(1+\gamma t) < \lambda(1+\gamma b)^2 \tag{29}$$
$$\textit{(stationary bound)} \quad \gamma L^2(1+\gamma t) < \lambda(1+\gamma b)^2 \tag{30}$$

Both the upper bound and stationary bound are satisfied by Assumption 1 (e). Further manipulating the lower bound, from Equation(28),

$$2\gamma\lambda + 2\gamma^2\lambda b - 2\gamma^2 L^2 < 1 + 2\gamma b + \gamma^2 b^2$$
$$\gamma^2(b^2 - 2\lambda b + 2L^2) + \gamma(2b - 2\lambda) + 1 > 0 \tag{31}$$

Solving the equality of Equation(31) (with the quadratic equation) gives us

$$\frac{(2\lambda - 2b) \pm \sqrt{(2b-2\lambda)^2 - 4(b^2 - 2\lambda b + 2L^2)}}{2(b^2 - 2\lambda b + 2L^2)}$$
$$= \frac{(2\lambda - 2b) \pm \sqrt{(4b^2 - 8\lambda b + 4\lambda^2) - 4b^2 + 8\lambda b - 8L^2}}{2(b^2 - 2\lambda b + 2L^2)}$$
$$= \frac{(2\lambda - 2b) \pm \sqrt{4\lambda^2 - 8L^2}}{2(b^2 - 2\lambda b + 2L^2)}$$
$$= \frac{(\lambda - b) \pm \sqrt{\lambda^2 - 2L^2}}{(b^2 - 2\lambda b + 2L^2)}$$

Recall that for a second-degree polynomial of the form $a_2 x^2 + a_1 x + 1$, the convexity is determined by a_2. Because $L \geq \lambda$ (a standard assumption), the discriminant $(\lambda^2 - 2L^2) < 0$ and thus there are no real roots. Looking at the convexity,

$$(b^2 - 2\lambda b + 2L^2) > (b^2 - 2\lambda b + \lambda^2) = (b - \lambda)^2 > 0$$

The strict inequality is because of the following. For all observed Fisher information matrices, (with p the dimesnion)

$$trace(\hat{\mathcal{I}}_n(\theta)) \geq b \Rightarrow \mathbb{E}trace(\hat{\mathcal{I}}_n(\theta)) \geq b \Rightarrow \lambda \cdot p \geq b$$

Thus for all $\gamma \in \mathbb{R}$ the lower bound represented by Equation(28) is satisfied. We have zero real roots and a convex function.

References

[1] Jyrki Kivinen, Manfred K Warmuth, and Babak Hassibi. The p-norm generalization of the lms algorithm for adaptive filtering. *Signal Processing, IEEE Transactions on*, 54(5):1782–1793, 2006.

[2] Brian Kulis and Peter L Bartlett. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 575–582, 2010.

[3] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[4] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2014.

[5] Panos Toulis, Jason Rennie, and Edoardo Airoldi. Statistical analysis of stochastic gradient methods for generalized linear models. In *31st International Conference on Machine Learning*, 2014.