Fast Rates of ERM and Stochastic Approximation: Adaptive to Error Bound Conditions

Mingrui Liu†, Xiaoxuan Zhang†, Lijun Zhang‡, Rong Jin♮, Tianbao Yang†

†Department of Computer Science, The University of Iowa ‡Nanjing University, Nanjing, China ♮Alibaba Group, Bellevue, WA 98004

Background and Main Contributions

Background: Error bound conditions (EBC) have recently received increasing attention in the field of optimization for developing faster convergence.

Contributions: Studied EBC in statistical learning setting.

1. Developed fast and optimistic rates of empirical risk minimization (ERM) under EBC for risk minimzation with Lipschitz continuous, smooth convex random functions.
2. Established fast rates of efficient stochastic approximation (SA) algorithm for risk minimization with Lipschitz continuous random functions, which requires only one pass of n samples and adapts to EBC.

Problem of Interest

1. Consider the **Risk Minimization** Problem:
$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[f(\mathbf{w},\mathbf{z})] \quad (1)$$
and more generally
$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[f(\mathbf{w},\mathbf{z})] + r(\mathbf{w}) \quad (2)$$
where $f(\cdot,\mathbf{z}) : \mathcal{W} \to \mathbb{R}$ is a random function depending on a random variable $\mathbf{z} \in \mathcal{Z}$ that follows a distribution $\mathbb{P}$, $r(\mathbf{w})$ is a lower semi-continuous convex function. $\mathcal{W} \subset \mathbb{R}^d$ is a convex and compact set (i.e., $\|\mathbf{w}\|_2 \le R$ for all $\mathbf{w} \in \mathcal{W}$), $\mathcal{W}_* = \arg\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w})$ denotes the optimal set and $P_* = \min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w})$ denotes the optimal risk.
2. $P(\mathbf{w})$ satisfies the error bound condition EBC(θ,α), i.e., for $\forall \mathbf{w} \in \mathcal{W}$,
$$\|\mathbf{w}-\mathbf{w}^*\|_2^2 \le \alpha(P(\mathbf{w}) - P(\mathbf{w}^*))^\theta, \quad (3)$$
where $\mathbf{w}^* = \arg\min_{\mathbf{u}\in\mathcal{W}_*}\|\mathbf{u}-\mathbf{w}\|_2$ denote an optimal solution closest to $\mathbf{w}$, $\mathcal{W}_*$ is the set containing all optimal solutions, $\theta \in (0,1]$ and $0 < \alpha < \infty$.

Other Conditions and Relationships to EBC

(Bernstein Condition) Let $\beta \in (0,1]$ and $B \ge 1$. Then $(f,\mathbb{P},\mathcal{W})$ satisfies the (β,B)-Bernstein condition if there exists a $\mathbf{w}_* \in \mathcal{W}$ such that for any $\mathbf{w} \in \mathcal{W}$
$$\mathbb{E}_{\mathbf{z}}[(f(\mathbf{w},\mathbf{z}) - f(\mathbf{w}_*,\mathbf{z}))^2] \le B(\mathbb{E}_{\mathbf{z}}[f(\mathbf{w},\mathbf{z}) - f(\mathbf{w}_*,\mathbf{z})])^\beta. \quad (4)$$

(v-Central Condition) Let $v : [0,\infty) \to [0,\infty)$ be a bounded, non-decreasing function satisfying $v(x) > 0$ for all $x > 0$. We say that $(f,\mathbb{P},\mathcal{W})$ satisfies the v-central condition if for all $\varepsilon \ge 0$, there exists $\mathbf{w}_* \in \mathcal{W}$ such that for any $\mathbf{w} \in \mathcal{W}$ the following holds with $\eta = v(\varepsilon)$.
$$\mathbb{E}_{\mathbf{z}\sim\mathbb{P}}\left[e^{\eta(f(\mathbf{w}_*,\mathbf{z})-f(\mathbf{w},\mathbf{z}))}\right] \le e^{\eta\varepsilon}. \quad (5)$$

EBC implies relaxed Berstein and v-central condition. Assume $f(\mathbf{w},\mathbf{z})$ is a G-Lipschitz continuous function w.r.t $\mathbf{w}$ for any $\mathbf{z} \in \mathcal{Z}$. For any $\mathbf{w} \in \mathcal{W}$, there exists $\mathbf{w}^* \in \mathcal{W}_*$ (which is actually the one closest to $\mathbf{w}$), such that
$$\mathbb{E}_{\mathbf{z}}[(f(\mathbf{w},\mathbf{z}) - f(\mathbf{w}^*,\mathbf{z}))^2] \le B(\mathbb{E}_{\mathbf{z}}[f(\mathbf{w},\mathbf{z}) - f(\mathbf{w}^*,\mathbf{z})])^\theta,$$
where $B = G^2\alpha$, and $\mathbb{E}_{\mathbf{z}\sim\mathbb{P}}\left[e^{\eta(f(\mathbf{w}^*,\mathbf{z})-f(\mathbf{w},\mathbf{z}))}\right] \le e^{\eta\varepsilon}$, where $\eta = v(\varepsilon) := c\varepsilon^{1-\theta} \wedge b$. Additionally, for any $\varepsilon > 0$ if $P(\mathbf{w}) - P(\mathbf{w}^*) \ge \varepsilon$, we have $\mathbb{E}_{\mathbf{z}\sim\mathbb{P}}\left[e^{v(\varepsilon)(f(\mathbf{w}^*,\mathbf{z})-f(\mathbf{w},\mathbf{z}))}\right] \le 1$, where $b > 0$ is any constant and $c = 1/(\alpha G^2\kappa(4GRb))$, where $\kappa(x) = (e^x - x - 1)/x^2$.

Remark: EBC does not require the existence of universal $\mathbf{w}^*$, which is required by original Bernstein and v-central condition.

Empirical Risk Minimization

Without loss of generality, we restrict our attention to (1) if $r(\mathbf{w})$ is a Lipschitz continuous convex function.

1. ERM for Lipschitz continuous random functions
Assume $f(\mathbf{w},\mathbf{z})$ is a G-Lipschitz continuous function w.r.t $\mathbf{w}$ for any $\mathbf{z} \in \mathcal{Z}$. If $r(\mathbf{w})$ is present, it can be absorbed into $f(\mathbf{w},\mathbf{z})$. It is notable that we do not assume $f(\mathbf{w},\mathbf{z})$ is convex in terms of $\mathbf{w}$ or any $\mathbf{z}$.

ERM for G-Lipschitz continuous random functions. For any $n \ge aC$, with probability at least $1-\delta$ we have
$$P(\widehat{\mathbf{w}}) - P_* \le O\left(\frac{d\log n + \log(1/\delta)}{n}\right)^{\frac{1}{2-\theta}}, \quad (6)$$
where $a = 3(d\log(32GRn^{1/(2-\theta)}) + \log(1/\delta))/c + 1$ and $C > 0$ is some constant.

2. ERM for non-negative, Lipschitz continuous and smooth convex random functions
Besides the Lipschitz continuity, we further assume $f(\mathbf{w};\mathbf{z})$ is a non-negative and L-smooth convex function w.r.t $\mathbf{w}$ for any $z \in \mathcal{Z}$. It is notable that we do not assume that $r(\mathbf{w})$ is smooth.

ERM for G-Lipschitz continuous and L-smooth random functions. With probability at least $1-\delta$ we have
$$P(\widehat{\mathbf{w}}) - P_* \le O\left(\frac{d\log n + \log(1/\delta)}{n} + \left[\frac{(d\log n + \log(1/\delta))P_*}{n}\right]^{\frac{1}{2-\theta}}\right).$$
When $n \ge \Omega\left(\left(\alpha^{1/\theta} d\log n\right)^{2-\theta}\right)$, with probability at least $1-\delta$,
$$P(\widehat{\mathbf{w}}) - P_* \le O\left(\left[\frac{d\log n + \log(1/\delta)}{n}\right]^{\frac{2}{2-\theta}} + \left[\frac{(d\log n + \log(1/\delta))P_*}{n}\right]^{\frac{1}{2-\theta}}\right).$$

Efficient SA for Lipschitz Continuous Random Functions

Algorithm 1 SSG$(\mathbf{w}_1,\gamma,T,\mathcal{W})$

Require: $\mathbf{w}_1 \in \mathcal{W}$, $\gamma > 0$ and T
Ensure: $\widehat{\mathbf{w}}_T$
1: **for** $t = 1,\ldots,T$ **do**
2: $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \gamma g_t)$
3: **end for**
4: $\widehat{\mathbf{w}}_T = \frac{1}{T+1}\sum_{t=1}^{T+1}\mathbf{w}_t$
5: return $\widehat{\mathbf{w}}_T$

Algorithm 2 ASA$(\mathbf{w}_1,n,R_0)$

1: Set $R_0 = 2R$, $\widehat{\mathbf{w}}_0 = \mathbf{w}_1$, $m = \lfloor\frac{1}{2}\log_2\frac{2n}{\log_2 n}\rfloor - 1$, $n_0 = \lfloor n/m\rfloor$
2: **for** $k = 1,\ldots,m$ **do**
3: Set $\gamma_k = \frac{R_{k-1}}{G\sqrt{n_0+1}}$ and $R_k = R_{k-1}/2$
4: $\widehat{\mathbf{w}}_k =$ SSG$(\widehat{\mathbf{w}}_{k-1},\gamma_k,n_0,\mathcal{W} \cap \mathcal{B}(\widehat{\mathbf{w}}_{k-1},R_{k-1}))$
5: **end for**
6: return $\widehat{\mathbf{w}}_m$

ASA for G-Lipschitz continuous random functions. Suppose $\|\mathbf{w}_1 - \mathbf{w}^*\|_2 \le R_0$, where $\mathbf{w}^*$ is the closest optimal solution to $\mathbf{w}_1$. Define $\bar{\alpha} = \max(\alpha G^2,(R_0G)^{2-\theta})$. For $n \ge 100$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, we have
$$P(\widehat{\mathbf{w}}_m) - P_* \le O\left(\frac{\bar{\alpha}(\log(n)\log(\log(n)/\delta))}{n}\right)^{\frac{1}{2-\theta}}.$$

Applications

Quadratic Problems (QP):
$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbf{w}^\top\mathbb{E}_{\mathbf{z}}[A(\mathbf{z})]\mathbf{w} + \mathbf{w}^\top\mathbb{E}_{\mathbf{z}'}[(\mathbf{z}')] + c \quad (7)$$
where c is a constant. The random function can be taken as $f(\mathbf{w},\mathbf{z},\mathbf{z}') = \mathbf{w}^\top A(\mathbf{z})\mathbf{w} + \mathbf{w}^\top(\mathbf{z}') + c$.

Remark: If $\mathbb{E}_{\mathbf{z}}[A(\mathbf{z})]$ is a positive semi-definite matrix (not necessarily positive definite) and $\mathcal{W}$ is a bounded polyhedron, then the problem (7) satisfies EBC$(\theta = 1,\alpha)$.

Piecewise Linear Problems (PLP):
$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}[f(\mathbf{w},\mathbf{z})] \quad (8)$$
where $\mathbb{E}[f(\mathbf{w},\mathbf{z})]$ is a piecewise linear convex function and $\mathcal{W}$ is a bounded polyhedron.

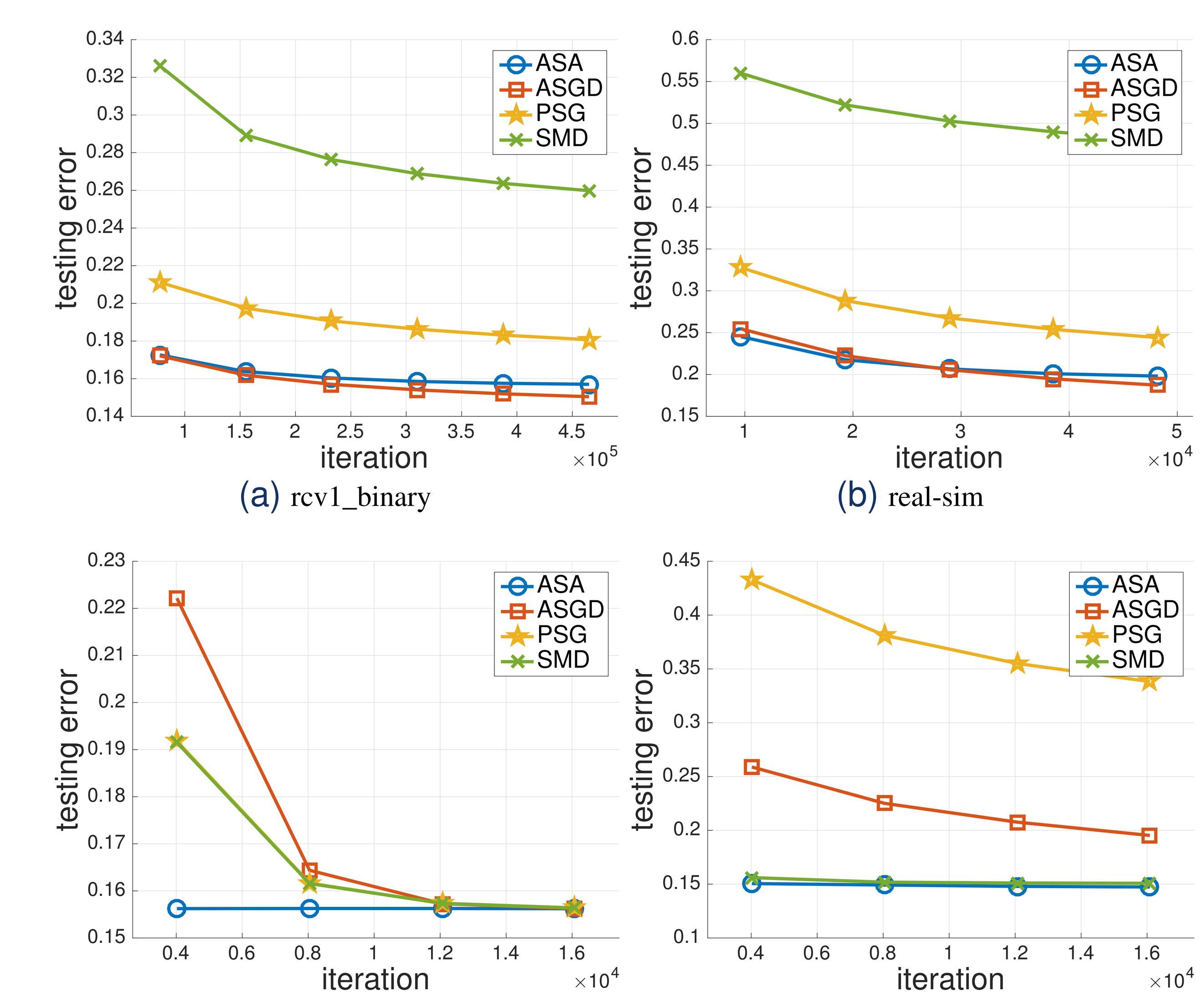
Remark: If $\mathbb{E}[f(\mathbf{w},\mathbf{z})]$ is piecewise linear and $\mathcal{W}$ is a bounded polyhedron, then the problem (8) satisfies EBC$(\theta = 1,\alpha)$.

Risk Minimization Problems over an ℓ_2 ball. Consider the following problem
$$\min_{\|\mathbf{w}\|_2\le B} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}}[f(\mathbf{w},\mathbf{z})] \quad (9)$$
Assuming that $P(\mathbf{w})$ is convex and $\min_{\mathbf{w}\in\mathbb{R}^d} P(\mathbf{w}) < \min_{\|\mathbf{w}\|_2\le B} P(\mathbf{w})$, we can show that EBC$(\theta = 1,\alpha)$ holds.

Risk Minimization with ℓ_1 Regularization Problems. For ℓ_1 regularized risk minimization:
$$\min_{\|\mathbf{w}\|_1\le B} P(\mathbf{w}) \triangleq \mathbb{E}[f(\mathbf{w};\mathbf{z})] + \lambda\|\mathbf{w}\|_1, \quad (10)$$

Remark: If the first component is quadratic as in (7) or is piecewise linear, then the problem (10) satisfies EBC$(\theta = 1,\alpha)$.

Experimental Results



(a) rcv1_binary (b) real-sim (c) E2006-tfidf (d) E2006-log1p

Figure: Testing Error vs Iteration of ASA and other baselines for SA when solving an ℓ_1 regularized expected square loss minimization problem