A Simple Analysis for Exp-concave Empirical Minimization with Arbitrary Convex Regularizer

Tianbao Yang*, Zhe Li*, Lijun Zhang♮

*The University of Iowa, ♮Nanjing University

Problem

Motivated by solving the stochastic composite optimization problem by Empirical Minimization:

$$\mathbf{w}_* = \arg\min_{\mathbf{w}\in\mathcal{W}} [P(\mathbf{w}) \triangleq \mathrm{E}_{\mathbf{z}\sim\mathcal{P}}[f(\mathbf{w},\mathbf{z})] + R(\mathbf{w})] \quad (1)$$

Study the convergence of the empirical minimizer of (1)

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left[P_n(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w},\mathbf{z}_i) + R(\mathbf{w}) \right] \quad (2)$$

- where $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are i.i.d samples from $\mathbb{P}$.

Goal: to establish the fast convergence rate of the empirical minimizer in terms of $P(\widehat{\mathbf{w}}) - P(\mathbf{w}_*)$.

Main Results

$P(\widehat{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \leq O(\frac{d\log n + d\log(1/\delta)}{n})$ and $F(\widetilde{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) \leq O(\frac{d\log n + d\log(1/\delta)}{n})$ both with high probability $1-\delta$.

Comparison with Related Works

The three recent studies [1, 2, 3] focus on establishing fast rates in terms of risk minimization without a regularizer

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) \triangleq \mathrm{E}_{\mathbf{z}}[f(\mathbf{w},\mathbf{z})], \quad (3)$$

- Koren & Levy [1] studied the convergence of a regularized empirical risk minimizer by

$$\widetilde{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left[\widehat{F}_n(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w},\mathbf{z}_i) + \frac{1}{n}g(\mathbf{w}) \right]. \quad (4)$$

- Mehta [2] targeted on the original risk mimization as (3)
- Gonen & Shalev-Shawarz [3] focused on the risk minimization with generalized linear model:

$$\min_{\mathbf{w}\in\mathcal{W}} L(\mathbf{w}) \triangleq \mathrm{E}_{(\mathbf{x},y)\sim D}[\phi_y(\mathbf{w}^\top\mathbf{x})], \quad (5)$$

Table 1: Difference of fast rates between our work and the related works [1, 2, 3]

Related Work	Ours
$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ [1, 2, 3]	$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) + R(\mathbf{w})$
*s-convex regularizer [1]	convex regularizer
Expectation [1, 3]	High probability[2]

*s-convex: strongly convex

- Our result is more general.

[1]. T. Koren and K. Y. Levy. Fast rates for exp- concave empirical risk minimization.

[2]. N. A. Mehta. Fast rate with high probability in exp-concave statistical learning.

[3]. A. Gonen and S. Shalev-Shwartz. Average stability is invariant to data preconditioning. implications to exp-concave empirical risk minimization.

Theoretical Results

Assumption 1:

- $\mathcal{W}$ is a closed and bounded convex set, i.e., there exists R such that $\|\mathbf{w}\|_2 \leq R$ for all $\mathbf{w}\in\mathcal{W}$.
- $f(\mathbf{w},\mathbf{z})$ is a G-Lipschitz continuous, L-smooth and β-exp-concave function of $\mathbf{w}\in\mathcal{W}$ for any $\mathbf{z}\in\mathcal{Z}$.
- $R(\mathbf{w})$ is a convex function.

Theorem 1

For the stochastic composite minimization problem (1), we consider the empirical minimizer $\widehat{\mathbf{w}}$ by solving (2). Under Assumption 1, with probability at least $1-\delta$, we have

$$P(\widehat{\mathbf{w}}) - P(\mathbf{w}_*) \leq O\left(\frac{d\log n}{n} + \frac{d\log(1/\delta)}{n\sigma}\right).$$

Remark 1: When $R(\mathbf{w}) = 0$, directly obtain a fast rate with high probability of the empirical risk minimizer for the exp-concave risk minimization.

Remark 2: Linear dependence on dimensionality d is unavoidable [4].

Theorem 2

For the risk minimization problem (3), we consider the regularized empirical risk minimizer $\widetilde{\mathbf{w}}$ by solving (4). Under Assumption 1 (i), (ii), and that $g(\mathbf{x})$ is bounded over $\mathcal{W}$ such that $\sup_{\mathbf{w},\mathbf{w}'\in\mathcal{W}} |g(\mathbf{w}) - g(\mathbf{w}')| \leq B$, with probability at least $1-\delta$, we have

$$F(\widetilde{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) \leq O\left(\frac{d\log n}{n} + \frac{d\log(1/\delta)}{n\sigma}\right).$$

Remark 1: Address the open problem raised in [1] about high probability bound for strongly regularized empirical risk minimizer.

Remark 2: Extend the fast rate to any regularized empirical risk minimizer as long as the regularizer is convex.

Analysis Technique

- Step 1: using the convexity of $P(\mathbf{w})$, the optimality condition of $\widehat{\mathbf{w}}$, and Cauchy-Schwarz inequality:

$$P(\widehat{\mathbf{w}}) - P(\mathbf{w}_*) \leq \|G(\widehat{\mathbf{w}},\mathbf{w}_*) - G_n(\widehat{\mathbf{w}},\mathbf{w}_*)\|_2 \|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 + \|\Delta_n(\mathbf{w}_*)\|_{H^{-1}} \|\widehat{\mathbf{w}} - \mathbf{w}_*\|_H$$

where $G(\mathbf{w},\mathbf{w}_*) = \nabla P(\mathbf{w}) - \nabla P(\mathbf{w}_*)$, $G_n(\mathbf{w},\mathbf{w}_*) = \nabla P_n(\mathbf{w}) - \nabla P_n(\mathbf{w}_*)$, $\Delta_n(\mathbf{w}) = \nabla P(\mathbf{w}) - \nabla P_n(\mathbf{w})$, H is local norm.

- Step 2: using concentration inequality [5], union bound and covering number of $\mathcal{W}$, with probability at least $1-\delta$,

$$\|G(\widehat{\mathbf{w}},\mathbf{w}_*) - G_n(\widehat{\mathbf{w}},\mathbf{w}_*)\|_2 \leq O(\frac{d\log n}{n}) + O(\sqrt{\frac{d(P(\widehat{\mathbf{w}}) - P(\mathbf{w}_*))}{n}}), \quad \|\Delta_n(\mathbf{w}_*)\|_{H^{-1}} \leq \frac{2G\log(2/\delta)}{n} + \sqrt{\frac{2\alpha d\log(2/\delta)}{n\sigma}}$$

- Step 3: using Young's inequality and do some linear algebra:

$$P(\widehat{\mathbf{w}}) - P(\mathbf{w}_*) \leq O\left(\frac{d\log n}{n} + \frac{d\log(1/\delta)}{n\sigma}\right).$$

[4]. V. Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back.

[5]. S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations.