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

Tianbao Yang*, Zhe Li*, Lijun Zhang†

*The University of Iowa, †Nanjing University

April 9, 2018

AISTATS 2018
Outline

1 Problem and Main Results

2 Comparison with Related Works

3 Theoretical Results

4 Analysis Technique

5 Conclusion

Fast Rate for Exp-concave ERM + Arbitrary Convex Regularizer
Motivated by solving the stochastic composite optimization problem by Empirical Minimization:

\[
\mathbf{w}_* = \arg \min_{\mathbf{w} \in \mathcal{W}} \left[ P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z} \sim \mathcal{P}}[f(\mathbf{w}, \mathbf{z})] + R(\mathbf{w}) \right]
\]  (1)
Motivated by solving the stochastic composite optimization problem by Empirical Minimization:

\[ w^*_\star = \arg \min_{w \in W} \left[ P(w) \triangleq E_{z \sim P}[f(w, z)] + R(w) \right] \]  \hspace{1cm} (1)
Motivated by solving the stochastic composite optimization problem by Empirical Minimization:  

\[
\mathbf{w}_* = \arg \min_{\mathbf{w} \in \mathcal{W}} \left[ \mathcal{P}(\mathbf{w}) \triangleq \mathbb{E}_{z \sim \mathcal{P}}[f(\mathbf{w}, z)] + R(\mathbf{w}) \right] \quad (1)
\]
Motivated by solving the stochastic composite optimization problem by Empirical Minimization:

\[
\mathbf{w}_* = \arg \min_{\mathbf{w} \in \mathcal{W}} \left[ P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z} \sim \mathcal{P}}[f(\mathbf{w}, \mathbf{z})] + R(\mathbf{w}) \right] 
\]  

\[
\mathcal{W} \text{ is compact and bounded convex set}
\]
Motivated by solving the stochastic composite optimization problem by Empirical Minimization:

$$w_* = \arg \min_{w \in \mathcal{W}} \left[ P(w) \triangleq \mathbb{E}_{z \sim \mathcal{P}}[f(w, z)] + R(w) \right]$$  \hspace{1cm} (1)

- $\mathcal{W}$ is compact and bounded convex set
- $f(w, z)$: smooth and $\beta$-exp-concave function of $w$ for any $z$, Lipschitz continuous over $\mathcal{W}$.
Motivated by solving the stochastic composite optimization problem by Empirical Minimization:

\[ w_* = \arg \min_{w \in \mathcal{W}} \mathbb{E}_{z \sim \mathcal{P}} [f(w, z)] + R(w) \]  

(1)

- \( \mathcal{W} \) is compact and bounded convex set
- \( f(w, z) \): smooth and \( \beta \)-exp-concave function of \( w \) for any \( z \), Lipschitz continuous over \( \mathcal{W} \).
- No assumption on \( R(w) \) except for convexity and boundness over \( \mathcal{W} \).
Study the convergence of the empirical minimizer of (1)

$$\hat{w} = \arg\min_{w \in \mathcal{W}} \left[ P_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + R(w) \right]$$

where $z_1, \ldots, z_n$ are i.i.d samples from $\mathbb{P}$. 
Problem

Study the convergence of the empirical minimizer of (1)

\[
\hat{w} = \arg \min_{w \in \mathcal{W}} \left[ P_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + R(w) \right]
\] (2)

where \(z_1, \ldots, z_n\) are i.i.d samples from \(P\).

Goal: to establish the fast convergence rate of the empirical minimizer in terms of \(P(\hat{w}) - P(w_*)\).
Main Theorem:

\[ \hat{w} = \arg \min_{w \in W} \left[ P_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + R(w) \right] \]

\[ P(\hat{w}) - \min_{w \in W} P(w) \leq O\left(\frac{d \log n + d \log(1/\delta)}{n}\right) \]

with high probability \(1 - \delta\), where \( P(w) \triangleq \mathbb{E}_{z \sim p}[f(w, z)] + R(w) \).
Main Results

Main Theorem:

\[
\hat{w} = \arg \min_{w \in \mathcal{W}} \left[ P_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + R(w) \right]
\]

\[
P(\hat{w}) - \min_{w \in \mathcal{W}} P(w) \leq O\left(\frac{d \log n + d \log(1/\delta)}{n}\right) \text{ with high probability } 1 - \delta, \text{ where } P(w) \triangleq E_{z \sim P}[f(w, z)] + R(w)
\]

Corollary:

\[
F(\tilde{w}) - \min_{w \in \mathcal{W}} F(w) \leq O\left(\frac{d \log n + d \log(1/\delta)}{n}\right) \text{ with high probability } 1 - \delta, \text{ where } F(w) \triangleq E_{z \sim P}[f(w, z)]
\]

\[
\tilde{w} = \arg \min_{w \in \mathcal{W}} \left[ \hat{F}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + \frac{1}{n} g(w) \right]
\]
1. Problem and Main Results

2. Comparison with Related Works

3. Theoretical Results

4. Analysis Technique

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

$$\min_{w \in \mathcal{W}} F(w) \triangleq \mathbb{E}_z[f(w, z)]$$  \hspace{1cm} (3)
The three recent studies [1, 2, 3] focus on establishing fast rates in terms of risk minimization without a regularizer.

\[
\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w}) \triangleq \mathbb{E}_z[f(\mathbf{w}, z)]
\]  

(3)

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

\[
\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathcal{W}} \left[ \hat{F}_n(\mathbf{w}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{w}, z_i) + \frac{1}{n} g(\mathbf{w}) \right]
\]

(4)
The three recent studies [1, 2, 3] focus on establishing fast rates in terms of risk minimization without a regularizer

$$\min_{w \in \mathcal{W}} F(w) \triangleq \mathbb{E}_z[f(w, z)]$$  \hspace{1cm} (3)

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

$$\tilde{w} = \arg \min_{w \in \mathcal{W}} \left[ \hat{F}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + \frac{1}{n} g(w) \right]$$  \hspace{1cm} (4)
Comparison with Related Works

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

\[
\min_{w \in \mathcal{W}} F(w) \triangleq \mathbb{E}_z[f(w, z)] \tag{3}
\]

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

\[
\hat{w} = \arg \min_{w \in \mathcal{W}} \left[ \hat{F}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + \frac{1}{n} g(w) \right] \tag{4}
\]

- Mehta [2] targeted on the original risk minimization as (3)
The three recent studies \cite{1, 2, 3} focus on establishing fast rates in terms of risk minimization \textit{without a regularizer}

\begin{equation}
\min_{w \in \mathcal{W}} F(w) \triangleq \mathbb{E}_z[f(w, z)]
\end{equation}

Koren & Levy \cite{1} studied the convergence of a regularized empirical risk minimizer by

\begin{equation}
\tilde{w} = \arg \min_{w \in \mathcal{W}} \left[ \hat{F}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^n f(w, z_i) + \frac{1}{n} g(w) \right]
\end{equation}

Mehta \cite{2} targeted on the original risk minimization as (3)

Gonen & Shalev-Shwartz \cite{3} focused on the risk minimization with generalized linear model:

\begin{equation}
\min_{w \in \mathcal{W}} L(w) \triangleq \mathbb{E}_{(x, y) \sim D}[\phi_y(w^\top x)]
\end{equation}
Comparison with Related Works

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

<table>
<thead>
<tr>
<th>Related Work</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \min_{w \in W} F(w) ) [1, 2, 3]</td>
<td>( \min_{w \in W} F(w) + R(w) )</td>
</tr>
<tr>
<td>*s-convex regularizer [1]</td>
<td>convex regularizer</td>
</tr>
<tr>
<td>Expectation [1, 3]</td>
<td>High probability [2]</td>
</tr>
</tbody>
</table>

*s-convex: strongly convex

- Our result is more general.


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

1. Problem and Main Results
2. Comparison with Related Works
3. Theoretical Results
4. Analysis Technique
5. Conclusion
Assumption 1

- $\mathcal{W}$ is a closed and bounded convex set, i.e., there exists $R$ such that $\|w\|_2 \leq R$ for all $w \in \mathcal{W}$. 
- $f(w, z)$ is a $G$-Lipschitz continuous, $L$-smooth and $\beta$-exp-concave function of $w \in \mathcal{W}$ for any $z \in Z$.
- $R(w)$ is a convex function.
Assumption 1

- $\mathcal{W}$ is a closed and bounded convex set, i.e., there exists $R$ such that $\|w\|_2 \leq R$ for all $w \in \mathcal{W}$.

- $f(w, z)$ is a $G$-Lipschitz continuous, $L$-smooth and $\beta$-exp-concave function of $w \in \mathcal{W}$ for any $z \in \mathcal{Z}$. 

Assumption 1

- \(\mathcal{W}\) is a closed and bounded convex set, i.e., there exists \(R\) such that \(\|w\|_2 \leq R\) for all \(w \in \mathcal{W}\).

- \(f(w, z)\) is a \(G\)-Lipschitz continuous, \(L\)-smooth and \(\beta\)-exp-concave function of \(w \in \mathcal{W}\) for any \(z \in \mathcal{Z}\).

- \(R(w)\) is a convex function.
Recall: \( P(w) \triangleq \mathbb{E}_{z \sim \mathcal{P}}[f(w, z)] + R(w) \)

**Theorem 1**

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

\[
P(\hat{w}) - P(w_*) \leq O \left( \frac{d \log n}{n} + \frac{d \log(1/\delta)}{n\sigma} \right).
\]
Theoretical Results

Recall: \( P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z} \sim \mathcal{P}}[f(\mathbf{w}, \mathbf{z})] + R(\mathbf{w}) \)

**Theorem 1**

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

\[
P(\hat{\mathbf{w}}) - P(\mathbf{w}^*) \leq O\left( \frac{d \log n}{n} + \frac{d \log(1/\delta)}{n\sigma} \right).
\]

**Remarks:**

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

- Linear dependence on dimensionality \( d \) is unavoidable [4].

Theoretical Results

Recall: \( F(w) \triangleq \mathbb{E}_{z \sim \mathcal{P}}[f(w, z)] \)
\[
\tilde{w} = \arg\min_{w \in \mathcal{W}} \left[ \hat{F}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + \frac{1}{n} g(w) \right]
\]
Theoretical Results

Recall: \( F(w) \triangleq \mathbb{E}_{z \sim \mathcal{P}}[f(w, z)] \)

\[ \tilde{w} = \arg\min_{w \in \mathcal{W}} \left[ \hat{F}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + \frac{1}{n} g(w) \right] \]

Theorem 2

Under Assumption 1 (i), (ii), and that \( g(x) \) is bounded over \( \mathcal{W} \) such that \( \sup_{w, w' \in \mathcal{W}} |g(w) - g(w')| \leq B \), with probability at least \( 1 - \delta \), we have

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

Remarks:
Theoretical Results

Recall: $F(w) \triangleq \mathbb{E}_{z \sim \mathcal{P}}[f(w, z)]$

$$\tilde{w} = \arg \min_{w \in \mathcal{W}} \left[ \hat{F}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + \frac{1}{n} g(w) \right]$$

**Theorem 2**

Under Assumption 1 (i), (ii), and that $g(x)$ is bounded over $\mathcal{W}$ such that $\sup_{w, w' \in \mathcal{W}} |g(w) - g(w')| \leq B$, with probability at least $1 - \delta$, we have

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

**Remarks:**

- Address the open problem raised in [1] about high probability bound for strongly regularized empirical risk minimizer.
Recall: $F(w) \triangleq \mathbb{E}_{z \sim P}[f(w, z)]$

$$\tilde{w} = \arg\min_{w \in \mathcal{W}} \left[ \hat{F}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w, z_i) + \frac{1}{n} g(w) \right]$$

### Theorem 2

Under Assumption 1 (i), (ii), and that $g(x)$ is bounded over $\mathcal{W}$ such that $\sup_{w, w' \in \mathcal{W}} |g(w) - g(w')| \leq B$, with probability at least $1 - \delta$, we have

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

### Remarks:

- Address the open problem raised in [1] about high probability bound for strongly regularized empirical risk minimizer.
- Extend the fast rate to any regularized empirical risk minimizer as long as the regularizer is convex.
1. Problem and Main Results
2. Comparison with Related Works
3. Theoretical Results
4. Analysis Technique
5. Conclusion
Step 1: by using the convexity of $P(w)$, the optimality condition of $\hat{w}$, and Cauchy-Schwarz inequality:

\[
P(\hat{w}) - P(w_*) \\
\leq \| G(\hat{w}, w_*) - G_n(\hat{w}, w_*) \|_2 \| \hat{w} - w_* \|_2 \\
+ \| \Delta_n(w_*) \|_{H^{-1}} \| \hat{w} - w_* \|_H
\]

where

\[
G(w, w_*) = \nabla P(w) - \nabla P(w_*), \\
G_n(w, w_*) = \nabla P_n(w) - \nabla P_n(w_*), \\
\Delta_n(w) = \nabla P(w) - \nabla P_n(w), \\
H = I + \frac{\sigma}{\alpha} E[\nabla f(w^*, z) \nabla f(w^*, z)^\top], \quad \sigma \text{ is parameter related to } \beta\text{-exp-concave and } \alpha \text{ is artificial parameter introduced.}
\]
Step 1: by using the convexity of $P(w)$, the optimality condition of $\hat{w}$, and Cauchy-Schwarz inequality:

$$
\begin{align*}
P(\hat{w}) - P(w_*) & \leq \|G(\hat{w}, w_*) - G_n(\hat{w}, w_*)\|_2 \|\hat{w} - w_*\|_2 \\
& + \|\Delta_n(w_*)\|_{H^{-1}} \|\hat{w} - w_*\|_H
\end{align*}
$$

where

$$
\begin{align*}
G(w, w_*) &= \nabla P(w) - \nabla P(w_*), \\
G_n(w, w_*) &= \nabla P_n(w) - \nabla P_n(w_*), \\
\Delta_n(w) &= \nabla P(w) - \nabla P_n(w), \\
H &= I + \frac{\sigma}{\alpha} \mathbb{E}[\nabla f(w^*, z) \nabla f(w^*, z)^\top],
\end{align*}
$$

$\sigma$ is parameter related to $\beta$-exp-concave and $\alpha$ is artificial parameter introduced.
Step 2: By using concentration inequality [5], union bound and covering number of $\mathcal{W}$, with probability at least $1 - \delta$,

$$\|G(\hat{w}, w_*) - G_n(\hat{w}, w_*)\|_2 \leq O\left(\frac{d \log n}{n}\right)$$

$$+ O\left(\sqrt{\frac{d(P(\hat{w}) - P(w_*))}{n}}\right)$$
Step 2: By using concentration inequality [5], union bound and covering number of $W$, with probability at least $1 - \delta$,

$$
\| G(\hat{w}, w^*) - G_n(\hat{w}, w^*) \|_2 \leq O\left( \frac{d \log \text{Variance}}{n} \right) + O\left( \sqrt{\frac{d (P(\hat{w}) - P(w^*))}{n}} \right)
$$
Step 2: By using concentration inequality [5], union bound and covering number of $\mathcal{W}$, with probability at least $1 - \delta$,

$$\| G(\hat{w}, w^*) - G_n(\hat{w}, w^*) \|_2 \leq O\left( \frac{d \log n}{n} \right) + O\left( \sqrt{\frac{d(P(\hat{w}) - P(w^*))}{n}} \right)$$

$$\| \Delta_n(w^*) \|_{H^{-1}} \leq \frac{2G \log(2/\delta)}{n} + \sqrt{\frac{2\alpha d \log(2/\delta)}{n\sigma}}$$
Step 2: By using concentration inequality [5], union bound and covering number of $\mathcal{W}$, with probability at least $1 - \delta$,

$$\| G(\hat{w}, w_*) - G_n(\hat{w}, w_*) \|_2 \leq O\left( \frac{d \log n}{n} \right)$$

$$+ O\left( \sqrt{d \left( P(\hat{w}) - P(w_*) \right)} \right)$$

$$\| \Delta_n(w_*) \|_{H^{-1}} \leq \frac{2G \log(2/\delta)}{n} + \sqrt{\frac{2\alpha d \log(2/\delta)}{n\sigma}}$$

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

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

with probability at least \( 1 - \delta \).
1. Problem and Main Results
2. Comparison with Related Works
3. Theoretical Results
4. Analysis Technique
5. Conclusion
Conclusion

- Developed a simple analysis of fast rates for empirical minimization with exp-concave loss functions and an arbitrary convex regularizer.
Developed a simple analysis of fast rates for empirical minimization with exp-concave loss functions and an arbitrary convex regularizer.

Exploited the covering number of a finite-dimensional bounded set and a concentration inequality of random vectors.
Developed a simple analysis of fast rates for empirical minimization with exp-concave loss functions and an arbitrary convex regularizer.

Exploited the covering number of a finite-dimensional bounded set and a concentration inequality of random vectors.

Induced an unified fast rate results for exp-concave empirical risk minimization without and with any convex regularizer.