# Fast Rates for Exp-concave Empirial Risk Minimization

Tomer Koren, Kfir Y. Levy
Presenter: Zhe Li

September 16, 2016

# High Level Idea

- Why could we use Regularized Empirical Risk Minimization?
  - Learning theory perspective
- If we could use Regularized Empirical Risk Minimization, how to solve?
  - Optimization algorithms

# Problem setting

Consider the problem of minimizing a stochastic objective

$$F(w) = \mathbb{E}[f(w, Z)]$$

- $w \in \mathcal{W} \subseteq \mathbb{R}^d$, a closed and convex domain $\mathcal{W}$

# Problem setting

Consider the problem of minimizing a stochastic objective

$$F(w) = \mathbb{E}[f(w, Z)]$$

- $w \in \mathcal{W} \subseteq \mathbb{R}^d$, a closed and convex domain $\mathcal{W}$
- random variable $Z$ distributed according to an unknow distribution over a parameter space $\mathcal{Z}$

# Problem setting

Consider the problem of minimizing a stochastic objective

$$F(w) = \mathbb{E}[f(w, Z)]$$

- $w \in \mathcal{W} \subseteq \mathbb{R}^d$, a closed and convex domain $\mathcal{W}$
- random variable $Z$ distributed according to an unknow distribution over a parameter space $\mathcal{Z}$
- given $n$ samples $z_1, \cdots, z_n$ of the random variable $Z$

# Problem setting

Consider the problem of minimizing a stochastic objective

$$F(w) = \mathbb{E}[f(w, Z)]$$

- $w \in \mathcal{W} \subseteq \mathbb{R}^d$, a closed and convex domain $\mathcal{W}$
- random variable $Z$ distributed according to an unknow distribution over a parameter space $\mathcal{Z}$
- given $n$ samples $z_1, \cdots, z_n$ of the random variable $Z$
- the goal: to produce an estimate $\hat{w} \in \mathcal{W}$ such that

$$\mathbb{E}[F(\hat{w})] - \min_w F(w)$$

is small.

# Assumptions

- $f(\cdot, z)$ is $\alpha$-exp-concave over the domain $\mathcal{W}$ for some $\alpha > 0$.
  - discuss later.
- $f(\cdot, z)$ is $\beta$-smooth over $\mathcal{W}$ with repect to Euclidean norm.
- $f(\cdot, z)$ is bounded over $\mathcal{W}$.

# Main results

How to construct an estimate $\hat{w}$?

- Based on the sample $z_1, \cdots, z_n$, construct

$$\hat{w} = \operatorname*{argmin}_{w \in \mathcal{W}} \hat{F}(w)$$

where

$$\hat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, z_i) + \frac{1}{n}R(w)$$

- $R(w) : \mathcal{W} \mapsto \mathbb{R}$: a regularizer, 1-strongly-convex w.r.t Euclidean norm. Assump that $|R(w) - R(w')| \leq B$ for all $w, w' \in W$ for constant $B > 0$

# Main results

**Theorem**

Let $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$ be a loss function defined over a closed and convex domain $\mathcal{W} \subseteq \mathbb{R}^d$, which $\alpha$-exp-concave,$\beta$-smooth and $C$ bounded w.r.t its first argument. Let $R : \mathcal{W} \mapsto \mathbb{R}$ be a 1-strongly-convex and $B$-bounded regularization function. Then for the regularized ERM estimate $\hat{w}$ based on an i.i.d samples $z_1, \cdots, \mathbf{z}_n$, the expected excess loss is bounded as

$$\mathbb{E}[F(\hat{w})] - \min_{w \in \mathcal{W}} F(\mathbf{w}) \leq \frac{24\beta d}{\alpha n} + \frac{100Cd}{n} + \frac{B}{n} = \mathcal{O}(d/n)$$

- Don't care about whatever optimazition algorithms
- Care about the learning framework

# Proof Technique

- Uniform Stability
- Average leave-one-out stablity
- Rademancher Complexity
- Local Rademancher Complexity

# Average leave-one-out stablity

- Define the empirical leave-one-out risk for each $i = 1, \cdots, n$

$$\hat{F}_i(w) = \frac{1}{n} \sum_{j \neq i} f(w, z_j) + \frac{1}{n} R(w)$$

- Let $\hat{w}_i = \underset{w \in \mathcal{W}}{\operatorname{argmin}} \hat{F}_i(w)$
- The average leave-one-out stability of $\hat{w}$ is defined as

$$\frac{1}{n} \sum_{i=1}^{n} (f(\hat{w}_i, z_i) - f(\hat{w}, z_i))$$

# Average leave-one-out stablity

**Theorem**

(Average leave-one-out stability). For any $z_1, \cdots, z_n \in \mathcal{Z}$ and for $\hat{w}_1, \cdots, \hat{w}_n$ and $\hat{w}$ as defined previously, we have

$$\frac{1}{n}\sum_{i=1}^{n}(f(\hat{w}_i, z_i) - f(\hat{w}, z_i)) \leq \frac{24\beta d}{\alpha n} + \frac{100Cd}{n}$$

# Proof of Main Theorem

- fix an arbitrary $w^* \in \mathcal{W}$, we have

$$F(w^*) + \frac{1}{n}R(w^*) = \mathbb{E}[\hat{F}(w^*)] \geq \mathbb{E}[\hat{F}(\hat{w})]$$

$$\Downarrow$$

$$\mathbb{E}[F(\hat{w}_n)] - F(w^*) \leq \mathbb{E}[F(\hat{w}_n) - \hat{F}(\hat{w})] + \frac{1}{n}R(w^*)$$

- since the random variable $\hat{w}_1, \cdots, \hat{w}_n$ have same distribution:

$$\mathbb{E}[F(\hat{w}_n)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[F(\hat{w}_i)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(\hat{w}_i, z_i)]$$

# Proof of Main Theorem

- $\mathbb{E}[\hat{F}(\hat{w})] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(\hat{w}, z_i)] + \frac{1}{n}\mathbb{E}[R(\hat{w})]$
- Combining the above inqualities,

$$\begin{aligned}
&\mathbb{E}[F(\hat{w}_n)] - F(w^*) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(\hat{w}_i, z_i) - f(\hat{w}, z_i)] + \frac{1}{n}\mathbb{E}[R(w^*) - R(\hat{w})] \\
&\leq \frac{24\beta d}{\alpha n} + \frac{100Cd}{n} + \frac{B}{n} = \mathcal{O}(\frac{d}{n})
\end{aligned}$$

using the average leave-one-out stability theorem and the assumption.

- $\square$

Proof of Average Leave-one-out Stability Theorem

# Local Strongly Convexity and Stability

**Definition**

(Local strong convexity). We say that a function $g : \mathcal{K} \mapsto \mathbb{R}$ is locally $\delta$-strongly convex over a domain $\mathcal{K} \subseteq \mathbb{R}^d$ at $x$ with respect to a norm $||\cdot||$,if

$$\forall y \in \mathcal{K}, g(y) \geq g(x) + \nabla g(x)(y-x) + \frac{\delta}{2}||y-x||^2$$

## Lemma

(Lemma 5). Let $g_1, g_2 : \mathcal{K} \mapsto \mathbb{R}$ be two convex functions defined over a closed and convex domain $\mathcal{K} \subseteq \mathbb{R}^d$, and let $x_1 \in \underset{x \in \mathcal{K}}{\operatorname{argmin}} g_1(x)$ and $x_2 \in \underset{x \in \mathcal{K}}{\operatorname{argmin}} g_2(x)$. Assume that $g_2$ is locally $\delta$-strongly convex at $x_1$ with repect to a norm $||\cdot||$. Then, for $h = g_2 - g_1$ we have

$$||x_2 - x_1|| \leq \frac{2}{\delta}||\nabla h(x_1)||^*$$

Futhermore, if $h$ is convex then

$$0 \leq h(x_1) - h(x_2) \leq \frac{2}{\delta}(||\nabla h(x_1)||^*)^2 \quad (1)$$

Some Definitions

- $f_i(\cdot) = f(\cdot, z_i)$ for all $i$, $h_i = \nabla f_i(\hat{w})$
- $H = \frac{1}{\delta} I_d + \sum_{i=1}^{n} h_i h_i^T$ and $H_i = \frac{1}{\delta} I_d + \sum_{j \neq i}^{n} h_i h_i^T$
- $||x||_M = \sqrt{x^T M x}$ denotes the norm induced by a positive definite matrix $M$, dual norm $||x||_M^* = \sqrt{x^T M^{-1} x}$

**Lemma**

(Lemma 6) For all $i = 1, \cdots, n$ it holds that

$$f_i(\hat{w}_i) - f_i(\hat{w}) \leq \frac{6\beta}{\delta} (||h_i||_{H_i}^*)^2$$

**Lemma**

(Lemma 8) Let $\mathcal{I} = \{i \in [n] : ||h_i||_H^* > \frac{1}{2}\}$. Then $|\mathcal{I}| \leq 2d$ and we have

$$\sum_{i \notin \mathcal{I}} (||h_i||_{H_i}^*)^2 \leq 2d$$

Lemma 6 + Lemma 8 $\Longrightarrow$ Stability Theorem

**Proof:**

- $\frac{1}{n} \sum_{i \in \mathcal{I}} (f_i(\hat{w}_i) - f_i(\hat{w})) \leq \frac{C|\mathcal{I}|}{n} \leq \frac{2Cd}{n}$
- $\frac{1}{n} \sum_{i \notin \mathcal{I}} (f_i(\hat{w}_i) - f_i(\hat{w})) \leq \frac{6\beta}{\delta n} \sum_{i \notin \mathcal{I}} (||h_i||_{H_i}^*)^2 \leq \frac{12\beta d}{\delta n}$
- summing up. $\square$

Continue to prove of Lemma 6 and Lemma 8?

# Proof of Lemma 8

**Lemma**

(Lemma 8) Let $\mathcal{I} = \{i \in [n] : ||h_i||_H^* > \frac{1}{2}\}$. Then $|\mathcal{I}| \leq 2d$ and we have

$$\sum_{i \notin \mathcal{I}} (||h_i||_{H_i}^*)^2 \leq 2d$$

**Proof**:

- $a_i = h_i^T H^{-1} h_i$ for $i = 1, \cdots, n$, $a_i > 0$.
- $\sum_i a_i \leq d$
- $|\mathcal{I}| \leq 2d$.
- $(||h_i||_{H_i}^*)^2 = h_i^T H_i^{-1} h_i^T = a_i + \frac{a_i^2}{1-a_i} \leq 2a_i$
- $\sum_{i \notin \mathcal{I}} (||h_i||_{H_i}^*)^2 \leq 2 \sum_{i \notin \mathcal{I}} a_i \leq \sum_i a_i = 2d$

# Proof of Lemma 6

**Lemma**

(Lemma 6) For all $i = 1, \cdots, n$ it holds that

$$f_i(\hat{w}_i) - f_i(\hat{w}) \leq \frac{6\beta}{\delta}(||h_i||^*_{H_i})^2$$

**Proof**:

- Using property of $\alpha$-exp-concave of function.
- Smoothness Assumption.
- Lemma 5.

# $\alpha$-exp-concave function

## Definition

The function $f(w)$ is $\alpha$-exp-concave over the domain $\mathcal{W}$ for some $\alpha > 0$, if that the function $\exp(-\alpha f(w))$ is concave over $\mathcal{W}$.

## Lemma

(Lemma 7) Let $f : \mathcal{K} \mapsto \mathbb{R}$ be an $\alpha$-exp-concave function over a convex domain $\mathcal{K} \subseteq \mathbb{R}^d$ such that $|f(x) - f(y)| \leq C$ for any $x, y \in \mathcal{K}$. Then for any $\delta \leq \frac{1}{2}\min\{\frac{1}{4C}, \alpha\}$, it holds that

$$\forall x, y \in \mathcal{K}, f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\delta}{2}(\nabla f(x)^T (y - x))^2$$

# Proof of Lemma 6

**Proof**:

- Let $g_1 = \hat{F}$ and $g_2 = \hat{F}_i$, $h_i = -\frac{1}{n} f_i$
  $\overset{Lemma7}{\Longrightarrow}$ $\hat{F}_i$ is locally $(\delta/n)$ strongly convex at $\hat{w}$ w.r.t $||\cdot||_{H_i}$
  $\overset{Lemma5}{\Longrightarrow}$ $||\hat{w}_i - \hat{w}||_{H_i} \leq \frac{2n}{\delta}||\nabla h(\hat{w})||^*_{H_i} = \frac{2}{\delta}||h_i||^*_{H_i}$
- $f_i$ is convex

$$
\begin{aligned}
f_i(\hat{w}_i) - f_i(\hat{w}) &\leq \nabla f_i(\hat{w}_i)^T(\hat{w}_i - \hat{w}) \\
&= \nabla f_i(\hat{w})^T(\hat{w}_i - \hat{w}) + (\nabla f_i(\hat{w}_i) - \nabla f_i(\hat{w}))^T(\hat{w}_i - \hat{w})
\end{aligned}
$$

# Proof of Lemma 6

- $\nabla f_i(\hat{w})^T(\hat{w}_i - \hat{w}) = h_i^T(\hat{w})^T(\hat{w}_i - \hat{w}) \leq ||h_i||_{H_i}^* \cdot ||\hat{w}_i - \hat{w}||_{H_i} \leq \frac{2}{\delta}(||h_i||_{H_i}^*)^2$
- $(\nabla f_i(\hat{w}_i) - \nabla f_i(\hat{w}))^T(\hat{w}_i - \hat{w}) \leq \beta||\hat{w}_i - \hat{w}||_2^2$
- $||\hat{w}_i - \hat{w}||_2^2 \leq \delta||\hat{w}_i - \hat{w}||_{H_i}^2 \leq \frac{4}{\delta}(||h_i||_{H_i}^*)^2$, by using $H_i \geq (1/\delta)I_d$.
- $\square$

# Open Question

- Is smoothness assumption necessary?
    - not necessary from online-batch convertion.
    - limition of the analysis?
- Excess risk with high probability?
    - Morkov's inequality?

# Open Question

- Is smoothness assumption necessary?
  - not necessary from online-batch convertion.
  - limition of the analysis?
- Excess risk with high probability?
  - Morkov's inequality? ✗