JMLR: Workshop and Conference Proceedings vol 40:1–16, 2015 28th Annual Conference on Learning Theory

Lower and Upper Bounds on the Generalization of Stochastic Exponentially Concave Optimization

Mehrdad Mahdavi MAHDAVI@UCHICAGO.EDU
Toyota Technological Institute at Chicago

Lijun Zhang ZHANGLJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University

Rong Jin RONGJIN@CSE.MSU.EDU
Michigan State University and Institute of Data Science and Technologies at Alibaba Group

Editors: Elad Hazan and Peter Grünwald

Abstract

In this paper we derive *high probability* lower and upper bounds on the excess risk of stochastic optimization of exponentially concave loss functions. Exponentially concave loss functions encompass several fundamental problems in machine learning such as squared loss in linear regression, logistic loss in classification, and negative logarithm loss in portfolio management. We demonstrate an $O(d \log T/T)$ upper bound on the excess risk of stochastic online Newton step algorithm, and an $O(d/T)$ lower bound on the excess risk of any stochastic optimization method for *squared loss*, indicating that the obtained upper bound is optimal up to a logarithmic factor. The analysis of upper bound is based on recent advances in concentration inequalities for bounding self-normalized martingales, which is interesting by its own right, and the proof technique used to achieve the lower bound is a probabilistic method and relies on an information-theoretic minimax analysis.

Keywords: stochastic optimization, exponentially concave losses, excess risk

1. Introduction

We study the generalization performance of stochastic optimization algorithms for exponentially concave losses. The problem of stochastic optimization is generally formulated as

$$\min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w}) \equiv \mathbb{E}\left[\ell(\mathbf{w};\xi)\right] = \int_{\Xi} \ell(\mathbf{w};\xi) d\mathrm{P}(\xi), \tag{1}$$

where domain $\mathcal{W} \subseteq \mathbb{R}^d$ is a closed convex set, ξ is a random variable taking values in Ξ, and P is a probability distribution over the instance space Ξ. The distribution of ξ may be unknown, but we assume only that we have access to a stochastic oracle that allows us to obtain independent and identically distributed (i.i.d.) samples $\xi_1, \xi_2, \ldots \in \Xi$ realized by underlying distribution P (Nemirovski et al., 2009). We study the performance of stochastic optimization algorithms characterized by the *excess risk* defined as:

$$\mathcal{L}(\mathbf{w}_T) - \min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w}),$$

where $\mathbf{w}_T$ is the solution obtained after receiving i.i.d samples $\xi_1, \ldots, \xi_T \in \Xi$.

Stochastic optimization has been well-studied when the objective function is convex or strongly convex. For general Lipschitz continuous convex functions, the stochastic gradient descent algorithm exhibits the optimal $O(1/\sqrt{T})$ rate of convergence (Nemirovski and Yudin, 1983; Nemirovski et al., 2009). For strongly convex functions, some variant of SGD (Juditsky and Nesterov, 2010; Rakhlin et al., 2012), such as the epoch gradient descent (Hazan and Kale, 2011), achieve an $O(1/T)$ convergence rate which is known to be minimax optimal (Agarwal et al., 2012). In terms of generalization bounds, an $O(1/T)$ high probability regret bound on the excess risk of strongly convex losses has been shown in (Kakade and Tewari, 2009) which is more appealing than the generalization of Lipschitz continuous cases.

In this study, we consider the scenario that the function $\ell(\cdot;\xi)$ is exponentially concave (abbr. exp-concave) (Cesa-Bianchi and Lugosi, 2006), a property which is stronger than convexity but weaker than strong convexity. In particular, the exp-concavity makes it possible to apply second-order methods and obtain theoretically superior convergence and/or regret bounds. This setting allows us to model many popular losses used in machine learning, such as the square loss in regression, logistic loss in classification (Hazan et al., 2014) and negative logarithm loss in portfolio management (Koren, 2013).

Despite its wide applicability, the stochastic optimization of exp-concave functions is not fully studied. In the online setting, (Hazan et al., 2007) have developed an Online Newton Step (ONS) algorithm that achieves a regret bound of $O(d \log T/T)$. Although a standard online-to-batch conversion (Cesa-Bianchi et al., 2004) of ONS yields an algorithm that attains an excess risk bound of $O(d \log T/T)$, this convergence rate only holds in *expectation* and does not precise the fluctuations of its risk. In fact, as it has been investigated in (Audibert, 2008), due to significant deviations of the estimator obtained by online-to-batch conversion from its expected performance, the high probability fast rates are not easily attainable in a closely related setting with exp-concave losses. Recently, (Mahdavi and Jin, 2014) have derived a high probability bound for a variant of ONS, but they impose a very strong assumption about the distribution. We believe there is considerable value in precisely characterizing the statistical risk of solutions in stochastic exp-concave optimization that holds with a high probability under standard assumptions such as boundedness and Lipschitzness.

To address the limitations mentioned above, we provide an in-depth analysis of ONS in stochastic setting. In particular, we demonstrate that ONS indeed achieves an $O(d \log T/T)$ excess risk bound that holds with a high probability. The only assumption that we make is the boundedness of the domain and the stochastic gradients, in contrast to the strong assumption made in (Mahdavi and Jin, 2014). Central to our analysis is a novel concentration inequality for bounding martingales, which is interesting by its own right. We also present an $\Omega(d/T)$ lower bound on the excess risk of any stochastic optimization method for squared loss as an instance of exp-concave losses. The proof of the lower bound is a probabilistic method that build off of an extension of information-theoretic Fano's inequality to stochastic optimization setting addressed here. In particular, we show that, for any stochastic optimization algorithm for minimizing squared loss, the linear dependency of excess risk on the dimensionality of data is unavoidable.

Organization The remainder of the paper is organized as follows. In Section 2 we describe the setting we consider, give the necessary background, and introduce the stochastic online Newton step algorithm. Also, in this section we state the main result on the lower and upper bounds on the generalization of exp-concave losses for stochastic ONS algorithm. We present the proof of upper bound in Section 3, and in Section 4 we provide an information-theoretic minimax analysis of the performance of any stochastic estimation method for exp-concave losses, though we defer few technical results to the appendices. We conclude in Section 5 with few problems as future research directions.

2. The Algorithm and Main Results

In this section we first provide definitions and assumptions that we need, and then introduce the stochastic online Newton step followed by stating the main theorems on the lower and upper bounds of the generalization of stochastic optimization of exp-concave functions.

We adopt the following notation throughout the paper. We use bold face lower case letters such as $\mathbf{w}$ to denote vectors and bold face upper case letters such as $\mathbf{A}$ to denote matrices. The ℓ_2 and weighted ℓ_2-norm of a vector $\mathbf{x} \in \mathbb{R}^d$ with respect to a positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ are denoted by $\|\mathbf{x}\|$ and $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$, respectively. The ℓ_0 of vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_0$. The dot product between two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\langle \mathbf{x}, \mathbf{y}\rangle$. The Frobenius norm of a matrix $\mathbf{A}$ is denoted by $\|\mathbf{A}\|_{\mathrm{F}}$. The transpose of a matrix $\mathbf{A}$ is denoted $\mathbf{A}^\top$. The identity matrix of dimension d is denoted by $\mathbf{I}_{d\times d}$, and if d is clear in the context, we will skip the subscript. The random variables are denoted by upper case letters such as X. The notation $\xi \sim \mathrm{P}$ indicates that the random variable ξ is drawn from the distribution P. For a random variable X measurable w.r.t. the randomness until round t, we use $\mathbb{E}_{t-1}[X]$ to denote its expectation conditioned on the randomness until round $t-1$. Finally, we use X_1^n to denote the sequence of random variables $X_1, X_2, \cdots, X_n$.

2.1. Definitions and Assumptions

The definition of exp-concave function is given below.

Definition 1 *(Cesa-Bianchi and Lugosi, 2006) A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is α-exp-concave over the convex domain $\mathcal{W} \subseteq \mathbb{R}^d$ if $\exp(-\alpha f(\cdot))$ is concave over $\mathcal{W}$.*

A nice property of exp-concave functions is that they can be approximated up to the second order by the their gradients, as indicated by the following lemma.

Lemma 1 *(Hazan et al., 2007, Lemma 3) For a function $f : \mathcal{W} \mapsto \mathbb{R}$, where $\mathcal{W}$ has diameter D, such that $\forall \mathbf{w} \in \mathcal{W}$, $\|\nabla f(\mathbf{w})\| \leq G$ and $\exp(-\alpha f(\cdot))$ is concave, the following holds for $\beta \leq \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\}$ and $\forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$:*

$$f(\mathbf{w}_1) \geq f(\mathbf{w}_2) + (\mathbf{w}_1 - \mathbf{w}_2)^\top \nabla f(\mathbf{w}_2) + \frac{\beta}{2}(\mathbf{w}_1 - \mathbf{w}_2)^\top \left[\nabla f(\mathbf{w}_2)\nabla f(\mathbf{w}_2)^\top\right](\mathbf{w}_1 - \mathbf{w}_2).$$

In order to apply the above lemma, we make the following assumptions about the stochastic optimization problem in (1).

(I) **Exp-cancavity** For all $\xi \in \Xi$, $\ell(\cdot;\xi)$ is α-exp-concave over domain $\mathcal{W}$.

(II) **Boundedness** The domain $\mathcal{W}$ is bounded: $\mathcal{W}$ has diameter D, i.e.,

$$\|\mathbf{w}_1 - \mathbf{w}_2\|_2 \leq D, \ \forall\ \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}. \tag{2}$$

(III) **Lipschitzness** The stochastic gradient is bounded:

$$\|\nabla \ell(\mathbf{w};\xi)\|_2 \leq G, \ \forall\ \mathbf{w} \in \mathcal{W}, \xi \in \Xi. \tag{3}$$

We note that in learning problem such as regression where $\xi = (\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, the condition of having a non-degenerate covariance for samples that each $\mathbf{x}$, the expected loss is strongly convex as well as exp-concave, and thus the excess risk bound could be $O(1/T)$. However, in this case the parameter of strong convexity is generally unknown, which makes it difficult to utilize existing algorithms. Furthermore, in high-dimensional setting, it is very likely that the covariance matrix degenerates, leading to a exp-concave loss rather than a strongly convex loss.

Algorithm 1 Stochastic Online Newton Step

input: Parameter λ
Initialize $\mathbf{w}_1$ as an arbitrary point in $\mathcal{W}$, and $\mathbf{Z}_1 = \lambda \mathbf{I}$
for $t = 1, \ldots, T$ **do**
  Receive $\xi_t \in \Xi$
  Calculate $\mathbf{Z}_{t+1}$ according to (5)
  Solve (4) to get $\mathbf{w}_{t+1}$
end for
return $\widehat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

2.2. Stochastic Online Newton Step Algorithm

The online learning approach is summarized in Algorithm 1, which is a combination of the online Newton step (ONS) (Hazan et al., 2007) with the online-to-batch conversion (Cesa-Bianchi et al., 2004).

In the beginning, we initialize the algorithm by setting $\mathbf{w}_1$ to be any point in $\mathcal{W}$ and $\mathbf{Z}_1 = \lambda \mathbf{I}$, where λ is a parameter that is introduced to ensure $\mathbf{Z}_t$ is invertible for all $t \geq 1$. At the tth round, a random sample ξ_t is received. To simplify the presentation, we define the instantaneous loss at round t on random sample ξ_t by $\ell_t(\mathbf{w}) = \ell(\mathbf{w}, \xi_t)$. Given the current solution $\mathbf{w}_t$, and the instantaneous loss $\ell_t(\cdot)$, the next solution $\mathbf{w}_{t+1}$ is obtained by solving the following convex optimization problem

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|^2_{\mathbf{Z}_{t+1}} + \frac{3}{\beta}(\mathbf{w}-\mathbf{w}_t)^\top \nabla \ell_t(\mathbf{w}_t) \tag{4}$$

where

$$\mathbf{Z}_{t+1} = \mathbf{Z}_t + \nabla \ell_t(\mathbf{w}_t)\nabla \ell_t(\mathbf{w}_t)^\top, \tag{5}$$

where $\beta > 0$ is the constant defined in Lemma 1. The matrix $\mathbf{Z}_{t+1}$ is an approximation of the Hessian matrix, and thus the updating rule is an analogue of the Newton–Raphson method. This is reason why it is refereed to as online Newton step (Hazan et al., 2007). In the last step, we return the average of all intermediate solutions, a standard operation which is dubbed as online-to-batch conversion (Cesa-Bianchi et al., 2004).

Denote the the minimizer of $\mathcal{L}(\mathbf{w})$ within domain $\mathcal{W}$ by $\mathbf{w}_*$, i.e., $\mathbf{w}_* = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w})$. Our first and main result is the following upper bound on the attainable excess risk by the stochastic ONS algorithm which holds in high probability.

Theorem 1 *Let $\widehat{\mathbf{w}}$ be the solution returned by Algorithm 1 after observing T random realizations of the loss function where each individual loss is β exp-concave. Then, with probability at least $1-2\delta$, we have*

$$\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) \leq O\left(\frac{d\left(\log T + \log\left(\frac{1}{\delta}\right)\right)}{\beta T}\right).$$

The next main result of this paper is the following lower bound on the excess risk of any stochastic exp-concave optimization method when utilize to minimize the square loss in regression problem.

Theorem 2 *Assume $d > 4$ and satisfy the condition $2^{d/4+1} \geq T(d-2)$. Then there exists there is a distribution over individual instances such that for any stochastic exp-concave optimization* *algorithm $\mathcal{A}$ applied to the least squared problem, for any solution $\widehat{\mathbf{w}}$ returned by the $\mathcal{A}$ after making T calls to the stochastic oracle, it holds that*

$$\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) = \Omega\left(\frac{d}{T}\right).$$

A few comments on Theorems 1 and 2 are in place here. First, as indicated in above theorem, the excess risk is reduced at the rate of $O(d \log T/T)$, which is consistent with the regret bound for online optimizing of exp-concave functions (Hazan et al., 2007). Second, for square loss which is a special case of exp-concave functions, a lower bound of $\Omega(d/T)$ has been established (Shamir, 2014) which matches the bound obtained here. This also indicates that the $O(d \log T/T)$ risk bound is optimal up to a logarithmic factor. We note that the linear dependence on d is in general unavoidable as stated in Theorem 2.

3. Generalization of Stochastic Online Newton Step

Our definitions and assumptions in place, we show in this section that stochastic ONS method enjoys a high-probability $O(d \log T/T)$ generalization guarantee for exp-concave loss functions. In particular, we prove the following result on the excess risk of the stochastic ONS algorithm. The omitted proofs are deferred to the appendix.

Theorem 3 *Let $\widehat{\mathbf{w}}$ be the solution returned by Algorithm 1 after observing T random realizations of the loss function. With a probability at least $1 - 2\delta$, we have*

$$\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) \leq \frac{1}{T}\left[\frac{\lambda\beta}{6}D^2 + \frac{3d}{2\beta}\log\left(1 + \frac{TG^2}{\lambda d}\right) + \Gamma_1 \log\frac{\sqrt{2T+1}}{\delta} + \Gamma_2\sqrt{\log\frac{2T+1}{\delta^2}}\right]$$

where

$$\Gamma_1 = \frac{24}{\beta} + \frac{8\beta G^2 D^2}{3} \text{ and } \Gamma_2 = 2GD + \frac{\beta G^2 D^2}{3}.$$

The proof build off of a concentration inequality for the sum of martingale difference sequences and characteristics of exp-concave functions discussed before. Our starting point is the following concentration inequality that underlies the proof of upper bound.

Theorem 4 *Let $\{X_i : i \geq 1\}$ be a martingale difference with respect to the filtration $\mathfrak{F} = \{\mathcal{F}_n : n \geq 1\}$ and suppose $|X_i^2| < R$ for all $i \geq 1$. Then, for any $0 < \delta < 1$, $\alpha > 0$, with a probability at least $1 - \delta$,*

$$\left|\sum_{i=1}^{t} X_i\right| \leq \alpha\left(\sum_{i=1}^{t} X_i^2 + \sum_{i=1}^{t}\mathbb{E}[X_i^2|\mathcal{F}_{i-1}]\right) + \frac{1}{\alpha}\log\frac{\sqrt{2t+1}}{\delta} + \sqrt{2}R\sqrt{\log\frac{2t+1}{\delta^2}},\ \forall t > 0.$$

The theorem will be proved later based on tools from the self-normalized processes (de la Peña et al., 2004; de la Peña and Pang, 2009; Abbasi-yadkori et al., 2011). It is remarkable that the concentration result our proof relies on makes our proof more elegant. If Freedman's inequality is applied as (Kakade and Tewari, 2009), two additional steps are required: (i) utilizing the peeling process to decouple dependence and (ii) taking a union bound to make the result holds for all $t > 0$.

We then introduce one lemma that is devoted to analyze the property of the updating rule in (4).

Lemma 2 *Let* $\mathbf{M}$ *be a positive definite matrix, and*

$$\mathbf{y} = \arg\min_{\mathbf{w}\in\mathcal{W}} \eta\langle \mathbf{w}, \mathbf{g}\rangle + \frac{1}{2}\|\mathbf{w}-\mathbf{x}\|_{\mathbf{M}}^2$$

Then for all $\mathbf{w}\in\mathcal{W}$, *we have*

$$\langle \mathbf{x}-\mathbf{w}, \mathbf{g}\rangle \leq \frac{1}{2\eta}\left(\|\mathbf{x}-\mathbf{w}\|_{\mathbf{M}}^2 - \|\mathbf{y}-\mathbf{w}\|_{\mathbf{M}}^2\right) + \frac{\eta}{2}\|\mathbf{g}\|_{\mathbf{M}^{-1}}^2.$$

Proof See Appendix A.1 for the proof. ■

To prove the result stated in Theorem 3, first according to Lemma 2, and based on the updating rule of stochastic ONS method, we have

$$\langle \mathbf{w}_t - \mathbf{w}_*, \nabla \ell_t(\mathbf{w}_t)\rangle \leq \frac{\beta}{6}\|\mathbf{w}_t-\mathbf{w}_*\|_{\mathbf{Z}_{t+1}}^2 - \frac{\beta}{6}\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{\mathbf{Z}_{t+1}}^2 + \frac{3}{2\beta}\|\nabla \ell_t(\mathbf{w}_t)\|_{\mathbf{Z}_{t+1}^{-1}}^2. \tag{6}$$

From the property of exp-concave functions in Lemma 1, we have

$$\ell_t(\mathbf{w}_*) \geq \ell_t(\mathbf{w}_t) + (\mathbf{w}_*-\mathbf{w}_t)^\top \nabla \ell_t(\mathbf{w}_t) + \frac{\beta}{2}(\mathbf{w}_*-\mathbf{w}_t)^\top \nabla \ell_t(\mathbf{w}_t)\nabla \ell_t(\mathbf{w}_t)^\top(\mathbf{w}_*-\mathbf{w}_t)$$

Taking the conditional expectation of both sides and rearranging, we obtain

$$\begin{aligned}
&\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_*)\\
\leq&\langle \nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_t-\mathbf{w}_*\rangle - \frac{\beta}{2}\mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t-\mathbf{w}_*)^\top\nabla\ell_t(\mathbf{w}_t)\right|^2\right]\\
=&\langle \nabla\ell_t(\mathbf{w}_t), \mathbf{w}_t-\mathbf{w}_*\rangle - \frac{\beta}{6}\|\mathbf{w}_t-\mathbf{w}_*\|_{\mathbf{Z}_{t+1}}^2 + \frac{\beta}{6}\|\mathbf{w}_t-\mathbf{w}_*\|_{\mathbf{Z}_{t+1}}^2\\
&- \frac{\beta}{2}\mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t-\mathbf{w}_*)^\top\nabla\ell_t(\mathbf{w}_t)\right|^2\right] + \langle \nabla\mathcal{L}(\mathbf{w}_t)-\nabla\ell_t(\mathbf{w}_t), \mathbf{w}_t-\mathbf{w}_*\rangle\\
\overset{(6)}{\leq}&-\frac{\beta}{6}\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{\mathbf{Z}_{t+1}}^2 + \frac{3}{2\beta}\|\nabla\ell_t(\mathbf{w}_t)\|_{\mathbf{Z}_{t+1}^{-1}}^2 + \frac{\beta}{6}\|\mathbf{w}_t-\mathbf{w}_*\|_{\mathbf{Z}_{t+1}}^2\\
&- \frac{\beta}{2}\mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t-\mathbf{w}_*)^\top\nabla\ell_t(\mathbf{w}_t)\right|^2\right] + \langle \nabla\mathcal{L}(\mathbf{w}_t)-\nabla\ell_t(\mathbf{w}_t), \mathbf{w}_t-\mathbf{w}_*\rangle\\
\overset{(5)}{=}&\frac{\beta}{6}\|\mathbf{w}_t-\mathbf{w}_*\|_{\mathbf{Z}_t}^2 - \frac{\beta}{6}\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{\mathbf{Z}_{t+1}}^2 + \frac{3}{2\beta}\|\nabla\ell_t(\mathbf{w}_t)\|_{\mathbf{Z}_{t+1}^{-1}}^2 + \frac{\beta}{6}\left|(\mathbf{w}_t-\mathbf{w}_*)^\top\nabla\ell_t(\mathbf{w}_t)\right|^2\\
&- \frac{\beta}{2}\mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t-\mathbf{w}_*)^\top\nabla\ell_t(\mathbf{w}_t)\right|^2\right] + \langle \nabla\mathcal{L}(\mathbf{w}_t)-\nabla\ell_t(\mathbf{w}_t), \mathbf{w}_t-\mathbf{w}_*\rangle.
\end{aligned}$$

By adding the above inequalities over all $t=1,\ldots,T$, we have

$$\begin{aligned}
&\sum_{t=1}^T \mathcal{L}(\mathbf{w}_t) - T\mathcal{L}(\mathbf{w}_*)\\
\leq&\frac{\lambda\beta}{6}D^2 + \frac{3}{2\beta}\underbrace{\sum_{t=1}^T\|\nabla\ell_t(\mathbf{w}_t)\|_{\mathbf{Z}_{t+1}^{-1}}^2}_{:=U_1^T} + \underbrace{\sum_{t=1}^T\langle\nabla\mathcal{L}(\mathbf{w}_t)-\nabla\ell_t(\mathbf{w}_t), \mathbf{w}_t-\mathbf{w}_*\rangle}_{:=U_2^T}\\
&+\frac{\beta}{6}\sum_{t=1}^T\left|(\mathbf{w}_t-\mathbf{w}_*)^\top\nabla\ell_t(\mathbf{w}_t)\right|^2 - \frac{\beta}{2}\sum_{t=1}^T\mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t-\mathbf{w}_*)^\top\nabla\ell_t(\mathbf{w}_t)\right|^2\right].
\end{aligned}\tag{7}$$

Now we proceed to bounding each individual terms U_1^T and U_2^T in the R.H.S of the above inequality on the excess risk. First, we discuss how to bound U_1^T. To this end, we have the following lemma.

Lemma 3

$$\sum_{t=1}^{T} \|\nabla \ell_t(\mathbf{w}_t)\|^2_{\mathbf{Z}_{t+1}^{-1}} \leq d \log\left(1 + \frac{TG^2}{\lambda d}\right). \tag{8}$$

Proof See Appendix A.2 for the proof. ■

Then, we utilize Theorem 4 to upper bound U_2^T, which is a summation of martingale difference sequences.

Lemma 4 *With a probability at least $1-\delta$, we have*

$$U_2^T \leq \frac{\beta}{12} \sum_{t=1}^{T} \left(\langle \nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2 + \mathbb{E}_{t-1}\left[\langle \nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2 \right] \right) + \Theta_T, \ \forall T > 0 \tag{9}$$

where

$$\Theta_T = \frac{24}{\beta} \log \frac{\sqrt{2T+1}}{\delta} + 2GD\sqrt{\log \frac{2T+1}{\delta^2}}.$$

Proof See Appendix A.3 for the proof. ■

Substituting (8) and (9) into (7), we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - T\mathcal{L}(\mathbf{w}_*) \leq \frac{\lambda\beta}{6} D^2 + \frac{3d}{2\beta} \log\left(1 + \frac{TG^2}{\lambda d}\right) + \Theta_T \\
&+ \frac{\beta}{12} \underbrace{\left(3\sum_{t=1}^{T} \left| (\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t) \right|^2 - 5\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[\left| (\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t) \right|^2 \right] \right)}_{:=U_3^T}
\end{aligned} \tag{10}$$

Next, we provide an upper for U_3^T in (10), which is also derived based on Theorem 4.

Lemma 5 *With a probability at least $1-\delta$, we have*

$$U_3^T \leq 4 \underbrace{\left(8G^2D^2 \log \frac{\sqrt{2T+1}}{\delta} + G^2D^2 \sqrt{\log \frac{2T+1}{\delta^2}} \right)}_{:=\Lambda_T}, \ \forall T > 0. \tag{11}$$

Plugging (11) in (10), we have

$$\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - T\mathcal{L}(\mathbf{w}_*) \leq \frac{\lambda\beta}{6} D^2 + \frac{3d}{2\beta} \log\left(1 + \frac{TG^2}{\lambda d}\right) + \Theta_T + \frac{\beta}{3}\Lambda_T$$

We complete the proof by noticing

$$\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - T\mathcal{L}(\mathbf{w}_*) \geq T\left(\mathcal{L}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t \right) - \mathcal{L}(\mathbf{w}_*) \right) = T\left(\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) \right),$$

which follows from the convexity of expected loss function and Jensen's inequality.

3.1. Proof of Lemma 5

We define a martingale difference sequences

$$Y_t = \left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2 - \mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2\right].$$

It is easy to verify that

$$|Y_t| \leq \max\left(\left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2, \mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2\right]\right) \leq G^2 D^2.$$

Applying Theorem 4 with $\alpha = 1/[8G^2D^2]$, with a probability at least $1 - \delta$, we have

$$\sum_{t=1}^T Y_t \leq \frac{1}{8G^2D^2}\left(\sum_{t=1}^T Y_t^2 + \sum_{t=1}^T \mathbb{E}_{t-1}[Y_t^2]\right) + \Lambda_T, \ \forall T > 0.$$

Following the proof of Lemma 4, we have

$$\begin{aligned}
\sum_{t=1}^T Y_t = & \sum_{t=1}^T \left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2 - \sum_{t=1}^T \mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2\right] \\
\leq & \frac{1}{4G^2D^2}\left(\sum_{t=1}^T \left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^4 + \sum_{t=1}^T \mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^4\right]\right) + \Lambda_T \\
\leq & \frac{1}{4}\left(\sum_{t=1}^T \left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2 + \sum_{t=1}^T \mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2\right]\right) + \Lambda_T
\end{aligned}$$

which implies

$$\frac{3}{4}\sum_{t=1}^T \left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2 - \frac{5}{4}\sum_{t=1}^T \mathbb{E}_{t-1}\left[\left|(\mathbf{w}_t - \mathbf{w}_*)^\top \nabla \ell_t(\mathbf{w}_t)\right|^2\right] \leq \Lambda_T.$$

3.2. Proof of Theorem 4

Before proving the concentration inequality in Theorem 4, we introduce two known results on the concentration of self-normalized processes from literature.

Theorem 5 *(de la Peña and Pang, 2009, Theorem 3.1) Let $\{X_i : i \geq 1\}$ be a martingale difference with respect to the filtration $\mathfrak{F} = \{\mathcal{F}_n : n \geq 1\}$ and suppose $\mathbb{E}[X_i^2] < \infty$ for all $i \geq 1$. Let τ be any stopping time with respect to the filtration $\mathfrak{F}$ and assume $\tau < \infty$ almost surely. Then for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^\tau X_i - \frac{\lambda^2}{2}\left(\sum_{i=1}^\tau X_i^2 + \sum_{i=1}^\tau \mathbb{E}[X_i^2|\mathcal{F}_{i-1}]\right)\right)\right] \leq 1.$$

Theorem 6 *(de la Peña et al., 2004, Theorem 2.1) Let A and B with $B > 0$ are two random variables such that*

$$\mathbb{E}\left[\exp\left(\lambda A - \frac{\lambda^2}{2}B^2\right)\right] \leq 1, \ \forall \lambda \in \mathbb{R}.$$

Then for all $y > 0$,

$$\mathbb{E}\left[\frac{y}{\sqrt{B^2+y^2}}\exp\left(\frac{A^2}{2(B^2+y^2)}\right)\right] \leq 1.$$

In the following, we define

$$A_\tau = \sum_{i=1}^{\tau} X_i \text{ and } B_\tau^2 = \sum_{i=1}^{\tau} X_i^2 + \sum_{i=1}^{\tau} \mathbb{E}[X_i^2|\mathcal{F}_{i-1}].$$

We are now in position to prove Theorem 4 inspired by the above two theorems. From above results we immediately get

$$\mathbb{E}\left[\frac{R}{\sqrt{B_\tau^2+R^2}}\exp\left(\frac{A_\tau^2}{2(B_\tau^2+R^2)}\right)\right] \leq 1. \tag{12}$$

By Markov's inequality, we have

$$\begin{aligned}
&\Pr\left[\frac{A_\tau^2}{2(B_\tau^2+R^2)} > \log\left(\sqrt{\frac{B_\tau^2+R^2}{R^2}}\frac{1}{\delta}\right)\right] \\
=&\Pr\left[\delta\frac{R}{\sqrt{B_\tau^2+R^2}}\exp\left(\frac{A_\tau^2}{2(B_\tau^2+R^2)}\right) > 1\right] \\
\leq&\mathbb{E}\left[\delta\frac{R}{\sqrt{B_\tau^2+R^2}}\exp\left(\frac{A_\tau^2}{2(B_\tau^2+R^2)}\right)\right] \overset{(12)}{\leq} \delta.
\end{aligned} \tag{13}$$

Then, we utilize the stopping time trick (Abbasi-yadkori et al., 2011) to derive a uniform bound that holds for all $t > 0$. Using the stopping time construction suggested by (Freedman, 1975), we define the bad event

$$E_t(\delta) = \left\{\omega : \frac{A_t^2}{2(B_t^2+R^2)} > \log\left(\sqrt{\frac{B_t^2+R^2}{R^2}}\frac{1}{\delta}\right)\right\}.$$

To bound the probability that $\bigcup_{t>0} E_t(\delta)$ happens, we define the stopping time $\tau(\omega) = \min\{t > 0 : \omega \in E_t(\delta)\}$ with the convention $\min\emptyset = \infty$. Evidently, $\bigcup\limits_{t>0} E_t(\delta) = \{\omega : \tau(\omega) < \infty\}$. Thus,

$$\begin{aligned}
&\Pr\left[\bigcup_{t>0} E_t(\delta)\right] = \Pr[\tau < \infty] \\
=&\Pr\left[\frac{A_\tau^2}{2(B_\tau^2+R^2)} > \log\left(\sqrt{\frac{B_\tau^2+R^2}{R^2}}\frac{1}{\delta}\right), \tau < \infty\right] \\
\leq&\Pr\left[\frac{A_\tau^2}{2(B_\tau^2+R^2)} > \log\left(\sqrt{\frac{B_\tau^2+R^2}{R^2}}\frac{1}{\delta}\right)\right] \overset{(13)}{\leq} \delta.
\end{aligned}$$

As a result, with a probability at least $1-\delta$, $\forall t > 0$, we have

$$A_t^2 \leq \log\left(\sqrt{\frac{B_t^2+R^2}{R^2}}\frac{1}{\delta}\right)2(B_t^2+R^2) \leq 2(B_t^2+R^2)\log\frac{\sqrt{2t+1}}{\delta}$$

which implies

$$|A_t| \leq \alpha B_t^2 + \frac{1}{\alpha}\log\frac{\sqrt{2t+1}}{\delta} + \sqrt{2}R\sqrt{\log\frac{2t+1}{\delta^2}},\ \forall\alpha > 0.$$

4. A Lower Bound On the Generalization of Stochastic Exp-Concave Optimization

We now show that for squared loss in regression problem, the minimax risk of any stochastic optimization method is $\Omega(d/T)$ as stated in Theorem 2. As a result, the stochastic Newton method algorithm achieves the almost optimal generalization error up to a logarithmic factor.

The proof of lower bound is a probabilistic method and utilizes results from (Duchi and Wainwright, 2013) to compute the lower bound of the minimax risk. Let V be a random variable taking values in a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \geq 2$. Here we consider a discrete set $\mathcal{V}$ and each element $v \in \mathcal{V}$ is associated with parameter $\theta_v := \theta\left(\mathrm{P}(\cdot|v)\right)$ that results in a distribution P. Let $\mathcal{P}$ be the distribution family created by the discrete set $\mathcal{V}$, and let $\rho : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ be a (semi)-metric on the parameter space, and let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ [1]. Then, we define the associated minimax *excess risk* for the family $\mathcal{P}$ as follows

$$\mathfrak{M}_n\left(\theta(\mathcal{P}), \Phi\circ\rho\right) = \inf_{\widehat{\theta}}\sup_{\mathrm{P}\in\mathcal{P}}\mathbb{E}_{\mathrm{P}}\left[\Phi\left(\rho\left(\widehat{\theta}(X_1^n), \theta(\mathrm{P})\right)\right)\right]$$

where $\theta(P)$ is the parameter for distribution P, X_1^n are n i.i.d sampled from distribution P, and $\widehat{\theta}(\cdot)$ is an estimator obtained using the sampled data points.

Our proof of lower bound is a probabilistic method and is based on the following result from (Duchi and Wainwright, 2013, Corollary 2) on the generalization of Fano's inequality:

Lemma 6 *Given V uniformly distributed over $\mathcal{V}$ with separation function $\delta(t)$, we have*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi\circ\rho) \geq \Phi\left(\frac{\delta(t)}{2}\right)\left(1 - \frac{I(X_1^n; V) + \log 2}{\log|\mathcal{V}| - \log N_t^{\max}}\right), \quad \forall t,$$

where $N_t^{\max} := \max_{v\in\mathcal{V}}|\{v' \in \mathcal{V}|\rho(v, v') \leq t\}|$ is the maximum neighborhood size at radius t and $\delta(t) := sup\{\delta|\rho(\theta_v, \theta_{v'}) \geq \delta$ for all $v, v' \in \mathcal{V}$ such that $\rho(v, v') \leq t\}$ is the separation of the set $\mathcal{V}$.

In our case, we are interested in the generalization error bound $\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*)$. To simplify our analysis and for the ease of exposition, we limit the analysis of lower bound to square loss, which is a special case of exponentially concave loss, and use a fixed design pattern $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{d\times T}$ (the generalization to other loss functions with linear classifiers is straightforward). We assume that the response Y are sampled from a Gaussian distribution, i.e., $Y \sim \mathcal{N}(\mathbf{X}^\top\mathbf{w}_*, \mathbf{I})$, where $\mathbf{w}_* \in \mathbb{R}^d$ is the parameter vector. It is easy to verify, under the above assumption, that for any estimator $\widehat{\mathbf{w}}$ the following equality on its excess risk holds:

$$\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) = \frac{1}{T}\|\mathbf{X}(\widehat{\mathbf{w}} - \mathbf{w}_*)\|^2.$$

1. For instance, for a univariate mean problem one can set $\rho(\theta - \theta') = |\theta - \theta'|$ and $\Phi(t) = t^2$, which leads to a risk in terms of mean-squared error.

Hence, we define the semi-metric as $\rho(\mathbf{w}, \mathbf{w}') = \frac{1}{\sqrt{T}}\|\mathbf{X}(\mathbf{w} - \mathbf{w}')\|^2$ and $\Phi(z) = z^2$. Using these notations, the minimax risk for the generalization error becomes as

$$\mathfrak{M}_T(\mathbf{X}) = \inf_{\widehat{\mathbf{w}}} \sup_{\mathrm{P} \in \mathcal{P}} \mathbb{E}_{\mathrm{P}} \left[\frac{1}{T} \left\| \mathbf{X}^\top (\widehat{\mathbf{w}}_Y - \mathbf{w}(\mathrm{P})) \right\|^2 \right],$$

where $\mathbf{w}(\mathrm{P})$ is used to represent the parameter vector for distribution P, and $\widehat{\mathbf{w}}_Y$ denotes the solution obtained based on $Y \in \mathbb{R}^T$ which is sampled from distribution P.

The following theorem bounds the minimax risk $\mathfrak{M}_T(\mathbf{X})$.

Theorem 7 *We have*

$$\mathfrak{M}_T(\mathbf{X}) > \frac{c\mu^2 d^2}{\|\mathbf{X}\|_F^2}$$

where

$$\mu^2 = \min \left\{ \frac{\|\mathbf{X}^\top \mathbf{z}\|_2^2}{dT} : \mathbf{z} \in \{-1, +1\}^d, \|\mathbf{z}\|_0 = \frac{d}{4} \right\}.$$

Proof To utilize Lemma 6, we introduce a discrete set $\mathcal{V} = \{\mathbf{v} \in \{-1, 1\}^d\}$, and define $\mathbf{w}_\mathbf{v} = \varepsilon \mathbf{v}$ for $\varepsilon > 0$. The family distribution $\mathcal{P}$ is then given by

$$\mathcal{P} = \{\mathcal{N}(\mathbf{X}\mathbf{w}, \mathbf{I}) | \mathbf{w} \in \mathcal{W}\}$$

where $\mathcal{W} = \{\mathbf{w} = \varepsilon \mathbf{v} : \mathbf{v} \in \mathcal{V}\}$. In our analysis, we set $t = d/4$. We have

$$\delta\left(\frac{d}{4}\right) \geq \min \left\{ \frac{2\varepsilon}{\sqrt{T}} \|\mathbf{X}^\top \mathbf{z}\|_2 : \mathbf{z} \in \{-1, +1\}^d, \|\mathbf{z}\|_0 = \frac{d}{4} \right\} = 2\sqrt{d}\mu\varepsilon$$

Using Lemma 6, we have

$$\mathfrak{M}_T(\mathbf{X}) > d\mu^2\varepsilon^2 \left(1 - \frac{I(V; Y) + \log 2}{\log |\mathcal{V}| - \log N_t^{\max}}\right).$$

In addition, we have

$$I(V; Y) \leq \frac{1}{|\mathcal{V}|} \sum_{\mathbf{u} \in \mathcal{V}} \sum_{\mathbf{v} \in \mathcal{V}} \mathrm{KL}\left(\mathcal{N}(\mathbf{X}\mathbf{w}_\mathbf{u}, \mathbf{I}) \| \mathcal{N}(\mathbf{X}\mathbf{w}_\mathbf{v}, \mathbf{I})\right) = \varepsilon^2 \|\mathbf{X}\|_\mathrm{F}^2$$

where $\mathrm{KL}(\cdot, \cdot)$ denotes the Kullback–Leibler divergence between two distributions (Cover and Thomas, 2012). We also have

$$\log |\mathcal{V}| - \log N_t^{\max} \geq cd$$

Combining the above results, we get

$$\mathfrak{M}_T(\mathbf{X}) > d\mu^2\varepsilon^2 \left(1 - \frac{\varepsilon^2 \|\mathbf{X}\|_\mathrm{F}^2}{cd}\right) \geq \frac{c\mu^2 d^2}{4\|\mathbf{X}\|_\mathrm{F}^2}$$

where the second inequality follows by setting $\epsilon^2 = cd/(2\|\mathbf{X}\|_\mathrm{F}^2)$. When d is sufficiently large, with high probability we have $\|\mathbf{X}\|_\mathrm{F}^2 = O(dT)$, and therefore we get $\mathfrak{M}_T(\mathbf{X}) > \frac{d}{T}$ as desired. ∎

To show the minimax risk is of $O(d/T)$, we need to construct a matrix $\mathbf{X}$ such that $\|\mathbf{X}\|_\mathrm{F}^2 = dT$ and μ^2 is sufficiently large constant, which is revealed by the following theorem.

Theorem 8 *Assume $d > 4$ and satisfy the condition $2^{d/4+1} \geq T(d-2)$. Then, there exists a matrix $\mathbf{X} \in \mathbb{R}^{d\times T}$, for which $\mu \geq 1/[2\sqrt{2}]$ holds.*

Proof We prove this theorem by a probabilistic argument. We construct $\mathbf{X}$ by sampling each entry $\mathbf{X}_{i,j}$ independently with $\Pr(\mathbf{X}_{i,j} = -1) = \Pr(\mathbf{X}_{i,j} = 1) = 1/2$. Evidently $\|\mathbf{X}\|_{\mathrm{F}}^2 = dT$. The proof of lower bound for μ is as follows. First, we establish a lower bound on the probability that $\|\mathbf{X}^\top \mathbf{z}\|_2^2/dT \geq 1/8$ holds for all $\Omega \equiv \{\mathbf{z} \in \{-1,+1\}^d, \|\mathbf{z}\|_0 = \frac{d}{4}\}$. Then, we estimate the expected number of such matrices $\mathbf{X} \in \{-1,+1\}^{d\times T}$ that satisfies the above inequality. This immediately implies that there exists at least a matrix $\mathbf{X}$ such that the above inequality holds.

Fix a matrix $\mathbf{X} \in \{-1,+1\}^{d\times T}$. Our goal is to lower bound the quantity $\mathrm{P}\left(\|\mathbf{X}^\top \mathbf{z}\|_2^2/dT \geq 1/8\right)$. For a column vector $\mathbf{x}_i$ in $\mathbf{X}$, we have $\mathbb{E}\left[|\mathbf{x}_i^\top \mathbf{z}|^2\right] = \frac{d}{4}$. Define

$$\delta = \mathrm{P}\left(|\mathbf{x}_i^\top \mathbf{z}| > \frac{\sqrt{d}}{2\sqrt{2}}\right).$$

Since $|\mathbf{x}_i^\top \mathbf{z}| \leq d/4$, we have $(1-\delta)\frac{d}{8} + \delta\frac{d^2}{16} \geq \frac{d}{4}$, and therefore $\delta > \frac{2}{d-2}$. Since $|\Omega| = \begin{pmatrix} d \\ d/4 \end{pmatrix}$, and we have T vectors in $\mathbf{X}$, hence, the probability for $\mu^2 \geq 1/8$ is $\delta/[T|\Omega|]$. Therefore by

$$\frac{2^d\delta}{T|\Omega|} > \frac{2^{d/4}\delta}{T} = \frac{2^{d/4+1}}{T(d-2)} \geq 1,$$

we guarantee to find a matrix $\mathbf{X}$ with $\mu \geq 1/[2\sqrt{2}]$. ■

5. Conclusions and Future Work

In this paper, we derived lower and upper bounds on the excess risk of exp-concave loss functions in stochastic setting. We derived an $O(d\log T/T)$ upper bound on the excess risk of a stochastic version of Online Newton Step algorithm. We presented a novel concentration inequality for martingales on which the proof of excess risk bound relies. For the regression problem with squared loss, we proved an $\Omega(d/T)$ lower bound using tools from information-theoretic analysis of minimax bounds. In particular, the proof of the lower bound was a probabilistic method using a simple generalization of Fano's inequality for minimax risk analysis. We believe that the obtained bounds on the generalization performance of exp-concave loss functions deepen our understanding exp-cancave loss functions including square loss, logistic loss, and logarithmic loss in stochastic optimization.

This work leaves few directions as future work. One open question to be addressed in the future is to investigate to see if the $\log T$ factor in the upper excess risk bound can be removed for exponentially concave loss functions by a more careful analysis. A solution to this question provides tighter bounds that matches the attainable minimax lower bound. Another open question that needs to be investigated in the future is to improve the dependence on d if we are after a sparse solution. In this regard, a question that is worthy of investigation is to see whether or not it is possible to obtain excess risk bounds stated in terms of sparsity of optimal solution rather than the dimension d.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful and insightful comments.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2008.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Victor H de la Peña and Guodong Pang. Exponential inequalities for self-normalized processes with applications. *Electronic Communications in Probability*, 14:372–381, 2009.

Victor H de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3):1902–1933, 2004.

John C Duchi and Martin J Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 197–209, 2014. URL http://jmlr.org/proceedings/papers/v35/hazan14a.html.

Anatoli Juditsky and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. Technical report, 2010.

Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2009.

Tomer Koren. Open problem: Fast stochastic exp-concave optimization. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 1073–1075, 2013.

Mehrdad Mahdavi and Rong Jin. Excess risk bounds for exponentially concave losses. *ArXiv e-prints*, arXiv:1401.4566, 2014.

A. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley & Sons Ltd, 1983.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied optimization*. Kluwer Academic Publishers, 2004.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456, 2012.

Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *ArXiv e-prints*, arXiv:1406.5143, 2014.

Appendix A. Omitted Proofs from the Analysis of Upper Bound

In this appendix we provide the proofs of few technical results omitted from the analysis of upper bound.

A.1. Proof of Lemma 2

Since $\mathbf{y}$ is the optimal solution to the following optimization problem

$$\mathbf{y} = \arg\min_{\mathbf{w}\in\mathcal{W}} \eta\langle \mathbf{w}, \mathbf{g}\rangle + \frac{1}{2}\|\mathbf{w}-\mathbf{x}\|_{\mathbf{M}}^2,$$

from the first order Karush-Kuhn-Tucker optimality condition (Nesterov, 2004; Boyd and Vandenberghe, 2004), we have

$$\langle \eta\mathbf{g} + \mathbf{M}(\mathbf{y}-\mathbf{x}), \mathbf{w}-\mathbf{y}\rangle \geq 0, \ \forall \mathbf{w}\in\mathcal{W}. \tag{14}$$

Then,

$$\begin{aligned}
&\|\mathbf{x}-\mathbf{w}\|_{\mathbf{M}}^2 - \|\mathbf{y}-\mathbf{w}\|_{\mathbf{M}}^2 \\
=&\mathbf{x}^\top \mathbf{M}\mathbf{x} - \mathbf{y}^\top \mathbf{M}\mathbf{y} + 2\langle \mathbf{M}(\mathbf{y}-\mathbf{x}), \mathbf{w}\rangle \\
\overset{(14)}{\geq}&\mathbf{x}^\top \mathbf{M}\mathbf{x} - \mathbf{y}^\top \mathbf{M}\mathbf{y} + 2\langle \mathbf{M}(\mathbf{y}-\mathbf{x}), \mathbf{y}\rangle - 2\langle \eta\mathbf{g}, \mathbf{w}-\mathbf{y}\rangle \\
=&\|\mathbf{y}-\mathbf{x}\|_{\mathbf{M}}^2 + 2\langle \eta\mathbf{g}, \mathbf{y}-\mathbf{x}+\mathbf{x}-\mathbf{w}\rangle \\
\geq&2\langle \eta\mathbf{g}, \mathbf{x}-\mathbf{w}\rangle + \min_{\mathbf{w}} 2\langle \eta\mathbf{g}, \mathbf{w}\rangle + \|\mathbf{w}\|_{\mathbf{M}}^2 \\
=&2\langle \eta\mathbf{g}, \mathbf{x}-\mathbf{w}\rangle - \eta^2\|\mathbf{g}\|_{\mathbf{M}^{-1}}^2.
\end{aligned}$$

where in the last inequality the scope of minimum over $\mathbf{w}$ is unbounded which is a clear lower bound.

By rearranging the terms we get

$$\langle \mathbf{x}-\mathbf{w}, \mathbf{g}\rangle \leq \frac{1}{2\eta}\left(\|\mathbf{x}-\mathbf{w}\|_{\mathbf{M}}^2 - \|\mathbf{y}-\mathbf{w}\|_{\mathbf{M}}^2\right) + \frac{\eta}{2}\|\mathbf{g}\|_{\mathbf{M}^{-1}}^2.$$

as desired.

A.2. Proof of Lemma 3

The analysis is almost identical to the proof of Lemma 11 in (Hazan et al., 2007). Following (Hazan et al., 2007, Lemma 12), we have

$$\|\nabla \ell_t(\mathbf{w}_t)\|_{\mathbf{Z}_{t+1}^{-1}}^2 \overset{(5)}{=} \langle \mathbf{Z}_{t+1}^{-1}, \mathbf{Z}_{t+1} - \mathbf{Z}_t\rangle \leq \log \frac{\det(\mathbf{Z}_{t+1})}{\det(\mathbf{Z}_t)},$$

and thus

$$\sum_{t=1}^T \|\nabla \ell_t(\mathbf{w}_t)\|_{\mathbf{Z}_{t+1}^{-1}}^2 \leq \sum_{t=1}^T \log \frac{\det(\mathbf{Z}_{t+1})}{\det(\mathbf{Z}_t)} = \log \frac{\det(\mathbf{Z}_{T+1})}{\det(\mathbf{Z}_1)}.$$

Recall that $\|\nabla \ell_t(\mathbf{w}_t)\| \leq G$. From (Abbasi-yadkori et al., 2011, Lemma 10), we have

$$\det(\mathbf{Z}_{t+1}) \leq \left(\lambda + \frac{TG^2}{d}\right)^d.$$

Since $\det(\mathbf{Z}_1) = \lambda^d$, we have

$$\log \frac{\det(\mathbf{Z}_{t+1})}{\det(\mathbf{Z}_1)} \leq d \log\left(1 + \frac{TG^2}{\lambda d}\right).$$

A.3. Proof of Lemma 4

Define $X_t = \langle \nabla \mathcal{L}(\mathbf{w}_t) - \nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_*\rangle$. From our assumption, we have

$$|X_t| \leq \|\nabla \mathcal{L}(\mathbf{w}_t) - \nabla \ell_t(\mathbf{w}_t)\|_2 \|\mathbf{w}_t - \mathbf{w}_*\|_2 \overset{(2),(3)}{\leq} 2GD.$$

Applying Theorem 4 with $\alpha = \beta/24$, with a probability at least $1-\delta$, we have

$$U_2^T \leq \frac{\beta}{24}\left(\sum_{t=1}^{T} X_t^2 + \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t^2]\right) + \frac{24}{\beta}\log\frac{\sqrt{2T+1}}{\delta} + 2GD\sqrt{\log\frac{2T+1}{\delta^2}},\ \forall T > 0. \quad (15)$$

Notice that

$$X_t^2 = \langle \nabla \mathcal{L}(\mathbf{w}_t) - \nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2 \leq 2\langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2 + 2\langle \nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2,$$
$$\mathbb{E}_{t-1}\left[X_t^2\right] = \mathbb{E}_{t-1}\left[\nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2\right] - \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2.$$

And thus using Jensen's inequality and some algebraic manipulation we get

$$\begin{aligned}
& X_t^2 + \mathbb{E}_{t-1}[X_t^2] \\
\leq & 2\langle \nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2 + \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2 + \mathbb{E}_{t-1}\left[\nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2\right] \\
\leq & 2\langle \nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2 + 2\mathbb{E}_{t-1}\left[\nabla \ell_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_* \rangle^2\right]
\end{aligned}$$

We complete the proof by plugging the above inequality in (15).