# A Generalization Bound for Online Variational Inference

**Badr-Eddine Chérief-Abdellatif** BADR.EDDINE.CHERIEF.ABDELLATIF@ENSAE.FR
*CREST, ENSAE, Institut Polytechnique de Paris*

**Pierre Alquier** PIERREALAIN.ALQUIER@RIKEN.JP
*RIKEN Center for AI Project, Tokyo, Japan*

**Mohammad Emtiyaz Khan** EMTIYAZ.KHAN@RIKEN.JP
*RIKEN Center for AI Project, Tokyo, Japan*

**Editors:** Wee Sun Lee and Taiji Suzuki

## Abstract

Bayesian inference provides an attractive online-learning framework to analyze sequential data, and offers generalization guarantees which hold even with model mismatch and adversaries. Unfortunately, exact Bayesian inference is rarely feasible in practice and approximation methods are usually employed, but do such methods preserve the generalization properties of Bayesian inference ? In this paper, we show that this is indeed the case for some variational inference (VI) algorithms. We consider a few existing online, tempered VI algorithms, as well as a new algorithm, and derive their generalization bounds. Our theoretical result relies on the convexity of the variational objective, but we argue that the result should hold more generally and present empirical evidence in support of this. Our work in this paper presents theoretical justifications in favor of online algorithms relying on approximate Bayesian methods.

**Keywords:** Bayesian inference, Variational inference, Online learning, Generalization bounds

## 1. Introduction

Bayesian methods, such as Kalman Filtering (Kalman, 1960), Hidden Markov Model (Baum and Petrie, 1966) and Particle Filtering (Doucet and Johansen, 2009), are popular methods to analyze sequential data. The posterior distribution provides a natural representation of the past information and can be updated sequentially using the Bayes rule whenever new data is available. Generalizations of Bayesian inference, such as those that *temper* the likelihood, offer good generalization guarantees (Banerjee, 2006; Audibert, 2009; Gerchinovitz, 2013). Such bounds hold even when the model is misspecified or when an adversary manipulates the stream of data. These generalization bounds are in fact very similar and sometimes even identical to the ones obtained by online learning methods commonly used in the optimization community (Cesa-Bianchi and Lugosi, 2006). The Bayesian principle offers a new perspective which can be used to advance online-learning methods used in areas such as convex optimization, machine learning, reinforcement learning, continual learning, and lifelong learning.

Unfortunately, exact Bayesian inference is computationally challenging in cases where the normalizing constant of the posterior distribution is a high-dimensional integral. Approximation methods, such as variational inference (VI) (Jordan et al., 1999) and expectation propagation (Minka, 2001), can dramatically reduce the computation cost and enable application of the Bayesian principle to large-scale problems. Despite concerns about their approximation error, these methods have extensively been applied to many machine-learning problems where they show satisfactory performance in practice (Blei and Lafferty, 2006; Hoffman et al., 2013; Kingma and Welling, 2013).

The practical success of such approximation methods points to the gap between the theory and practice. A few recent works have established generalization bounds of the approximation methods such as variational inference, but these are restricted to the batch or offline setting (Alquier and Ridgway, 2017; Bhattacharya et al., 2018; Zhang and Gao, 2017). Extending such results to the online setting, without making strong assumption about the model mismatch and adversaries, is the main focus of this paper.

We propose online version of variational inference with tempered likelihoods, and derive new generalization bound, which has very similar form to the bound of exact Bayesian inference. Unlike existing proof techniques, our proof extend to the case when approximations are used instead of the exact Bayesian update. Our derivation relies on the convexity of the variational objective. This covers a few important cases, but can be limiting. We argue that the generalization bound is likely to hold more generally, and present empirical evidence in support of these arguments. Our work takes a step towards establishing the generalization properties of online approximate Bayesian methods.

### 1.1. Related works

Variational inference is extremely popular in statistics and machine learning, yet its theoretical properties are not investigated until recently. Generalization bounds for generalized versions of variational approximations are derived in Alquier et al. (2016); Cottet and Alquier (2018). Similarly, Bernstein-von Mises' theorems for variational approximations in parametric models are proved in Wang and Blei (2018), while concentration of the posterior in general models is studied in Alquier and Ridgway (2017); Sheth and Khardon (2017); Bhattacharya et al. (2018); Zhang and Gao (2017); Chérief-Abdellatif and Alquier (2018); Chérief-Abdellatif (2019); Jaiswal et al. (2019). These works show that variational approximations does enjoy the same consistency properties as the posterior distribution under general conditions. All of these results however only apply to the batch setting and their extension to the online setting is not straightforward.

It is known that the Bayesian approach leads to good online predictions for a stream of data; see Banerjee (2006), and Cesa-Bianchi and Lugosi (2006); Audibert (2009); Gerchinovitz (2013) for generalized posteriors in machine learning. However, there are only a few attempts to study the online properties of variational inference, and the proofs used in Cesa-Bianchi and Lugosi (2006) cannot easily be extended to online variational inference.

Generalization bounds for online approximations of the posterior are studied in Guhaniyogi et al. (2013), but the algorithms analyzed there are different from the ones used in practice and the feasibility of these algorithms is not proven. Recently Nguyen et al. (2017a) give some results, but the order of magnitude of the bounds are not explicitly written and in many contexts it is not clear that the bound will even be small enough to ensure consistency. Even though stochastic/online versions of variational inference are known to give good results in practice (Sato, 2001; Hoffman et al., 2010; Wang et al., 2011; Hoffman et al., 2013; Khan and Lin, 2017; Nguyen et al., 2017b; Khan and Nielson, 2018; Khan et al., 2018; Zeno et al., 2018), existing works have not been able to derive theoretical results confirming their generalization properties. Our results fill this gap between theory and practice for some types of variational approximations obtained with specific types of online algorithms.

## 2. Generalization Properties of Bayesian Inference for Online Learning

Given a stream of data, the goal of online learning is to learn to make good decisions, estimations, or predictions on future data examples. The quality of such decisions is defined with a loss function $\ell(\mathcal{D}_t, \hat{\theta}_t)$, denoted by $\ell_t(\hat{\theta}_t)$ for brevity, where $\mathcal{D}_t$ is the data at time $t$ and $\hat{\theta}_t$ is a quantity computed using the past data, i.e., $\mathcal{D}_{1:(t-1)} := \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_{t-1}\}$. This definition of the loss includes popular supervised and unsupervised learning methods. For example, in maximum-likelihood training of a parameterized model $p_\theta$, $\hat{\theta}_t$ is the parameter estimate and the loss is $\ell_t(\theta) := -\log p_\theta(\mathcal{D}_t)$. Similarly, for a classification task with input-output pair $\mathcal{D}_t := (X_t, Y_t)$, the loss could be the hinge loss $\ell_t(\theta) = (1 - Y_t f_\theta(X_t))_+$ with a classifier $f_\theta$. In the whole paper, we assume that $\theta \mapsto \ell_t(\theta)$ is convex. By using losses $\ell_t$ until time $t$, our ultimate goal is to find a $\theta_t$ which is as close as possible to the minimizer $\theta^*$ of the generalization error $\mathcal{E}_*(\theta) = \mathbb{E}_{\mathcal{D}\sim P_*}[\ell(\mathcal{D}, \theta)]$ where $P_*$ is the true distribution of the data. We would want to do this without many strong assumptions such as assuming the data stream to be i.i.d., or the absence of adversaries.

Since $\mathcal{E}_*$ is unavailable at time $t$, to ensure the quality of $\hat{\theta}_t$, online-learning algorithms aim at minimizing the cumulative error $\sum_{i=1}^{t} \ell_i(\hat{\theta}_t)$ until time $t$. Many algorithms are known with bounds on the *regret* of the decision $\hat{\theta}_t$ , that is the gap in the cumulative error and the minimal cumulative error that could have been reached with a *fixed* parameter:

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{T} \ell_t(\theta). \tag{1}$$

Bounds on this quantity are known as *regret* bounds, e.g., see Cesa-Bianchi and Lugosi (2006); Bubeck (2011); Shalev-Shwartz (2012); Hazan (2016). Fortunately, bounding the regret also leads to upper bounds on the generalization gap, e.g., by using the average $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}_t$ we can bound the gap $\mathcal{E}_*(\bar{\theta}_T) - \mathcal{E}_*(\theta^*)$. Due to such properties, regret bounds are useful to study generalization properties of an online algorithm. Moreover, the bound holds with very little assumptions on the data and is valid when the data is not i.i.d. and even when it is corrupted by an adversary.

For online learning, Bayesian inference algorithms have good generalization properties, e.g., the following *tempered* posterior distribution introduced by Vovk (1990); Littlestone and Warmuth (1994) has a controlled regret:

$$p_t^\eta(\theta) := \frac{1}{\mathcal{Z}_t^\eta} \pi(\theta) e^{-\eta \sum_{i=1}^{t-1} \ell_t(\theta)} \tag{2}$$

where $\eta > 0$ is a learning rate, $\pi$ is a prior distribution, and $\mathcal{Z}_t^\eta$ is the normalizing constant of the posterior distribution. Each loss $\ell_t$ here can be interpreted as the log-likelihood of a data example $\mathcal{D}_t$. When the loss is indeed equal to $-\log p_\theta(\mathcal{D})$ and $\eta = 1$, the above

**Algorithm 1** Tempered Bayesian Inference, a.k.a Exponentially Weighted Aggregration

```
Input Learning rate $\eta > 0$, prior $\pi(\theta)$, $p_1^\eta \leftarrow \pi$.
For $t = 1, 2, 3, \dots,$
    1. $\hat{\theta}_t \leftarrow \mathbb{E}_{\theta \sim p_t^\eta}(\theta)$,
    2. Observe $\mathcal{D}_t$ to suffer a loss $\ell_t(\hat{\theta}_t)$.
    3. Update $p_{t+1}^\eta(\theta) \propto p_t^\eta(\theta) \exp\left[-\eta \ell_t(\theta)\right]$.
```

algorithm is equivalent to Bayesian inference whose generalization properties are usually established under the assumption of no model mismatch (e.g., see Ghosal and Van der Vaart (2017)). The tempered version $\eta < 1$ can be shown to generalize well even when the model is misspecified (Grünwald and Van Ommen, 2017) or when an adversary manipulates the stream of data. Such tempered versions have also been studied in depth in the machine-learning literature by using the PAC-Bayesian bounds (Shawe-Taylor and Williamson, 1997; McAllester, 1999; Catoni, 2007; Seldin and Tishby, 2010; Suzuki, 2012; Seldin et al., 2011; Cuong et al., 2013; Germain et al., 2016; Catoni and Giulini, 2017; Guedj, 2019; Tsuzuku et al., 2019).

In the online-learning literature, the regret bound of this algorithm has been studied extensively under a variety of names, e.g., algorithms such as multiplicative update, weighted majority algorithm, exponentially weighted aggregation (EWA) are all specific cases of tempered Bayesian inference. Algorithm 1 shows a pseudo-code for EWA which performs tempered Bayesian inference in an online fashion (Step 3 implements Equation (2)). Below, we state a theorem which shows an example of regret bound[1], proved in Theorem 4.6 in Audibert (2009) for the algorithm shown in Algorithm 1.

**Theorem 1** *Assuming that the loss is bounded, i.e.,* $0 \leq \ell_t(\theta) \leq B$, $\forall \mathcal{D}_t, \theta$, *the cumulative regret has the following upper bound when* $\hat{\theta}_t = \mathbb{E}_{\theta \sim p_t^\eta}[\theta]$ *is the posterior mean:*

$$\sum_{t=1}^T \ell_t(\hat{\theta}_t) \leq \inf_{p \in \mathcal{S}} \left\{ \mathbb{E}_{\theta \sim p} \left[ \sum_{t=1}^T \ell_t(\theta) \right] + \frac{\eta B^2 T}{8} + \frac{\mathcal{K}(p, \pi)}{\eta} \right\} \tag{3}$$

*where* $\mathcal{S}$ *is the set of all probability distributions over* $\Theta$ *and* $\mathcal{K}$ *is the Küllback-Leibler (KL) divergence.*

A proof is given in Appendix 6.5 for the sake of completeness.

The above regret bound is useful to derive explicit bounds in expectation on the generalization error $\mathcal{E}_*$ of an estimator that is defined as the average decision $\bar{\theta}_T := \sum_t \hat{\theta}_t / T$. For example, we can show that, when a classical prior mass condition[2] on the prior is satisfied

[1] In online-learning literature such results are usually stated for finite decision space, e.g., see similar results for EWA in Cesa-Bianchi and Lugosi (2006). The result above holds for a more general continuous setting but under a bounded loss.

[2] The exact condition is that the prior $\pi(\theta)$ has mass bigger than $\epsilon^d$ on an $\epsilon$-ball around $\theta^*$ for some $d$.

and when $\mathcal{D}_t$ are actually independent and identically distributed from $P_*$, the generalization error has the following bound:

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\mathcal{E}_*(\bar{\theta}_T)\right] \leq \mathcal{E}_*(\theta^*) + B\sqrt{\frac{d}{2T}\log\left(\frac{T}{d}\right)} \tag{4}$$

for some well-chosen $\eta \sim \sqrt{d/T}$ and $d > 0$ is a complexity measure of the parameter space (often the dimension). This bound shows that when $\mathcal{D}_t$ are i.i.d. from $P^*$ then Bayesian inference achieves generalization error at a rate $\sqrt{d/T}$. An exact statement and a complete proof are given in Theorem 6 Subsection 6.3 in the appendix. The proof is based on a technique called *online-to-batch* analysis. Similar bounds can be derived even for the cases when the model is misspecified and an adversary is present.

The regret bound derived in Theorem 1 assumes that $p_t^\eta$ is computed exactly, which is extremely challenging and many a times infeasible. The difficulty arises due to the computation of $\mathcal{Z}_t^\eta$ which is a high-dimensional integral when the space of $\theta$ is large. Approximate Bayesian inference methods approximate the integral by finding an approximation of $p_t^\eta$ in a restricted family of distributions $\mathcal{F} = \{q_\mu, \mu \in \mathcal{M}\}$, e.g., Gaussian distribution with $\mu$ being the mean and variance. Our focus in this paper is to derive bounds similar to Theorem 1 for approximate Bayesian inference methods.

Unfortunately, deriving similar bounds as Theorem 1 for approximate inference is not possible using existing proof techniques. This is because these techniques do not work when $p_t^\eta$ and $\mathcal{S}$ in (3) are replaced by $q_{\mu_t}$ and $\mathcal{M}$ respectively. As shown in Appendix 6.5, these proofs rely on cancellation of many terms in a telescoping sum. This cancellation does not take place when an approximation is used instead, and the error accumulates making the regret bound practically useless. In this paper, we solve this problem using a different proof for tempered, online variational inference algorithms discussed in the next section.

## 3. Online Variational Inference

In this section, we introduce approximate Bayesian inference methods that can obtain tractable approximations in an online fashion. The methods available in the approximate inference literature are not always suitable for our purpose. Therefore, we present modifications of those methods that lead to feasible online variants of the Bayesian update shown in (2). To simplify the notation, we will denote the expectation of the loss under an approximation $q_\mu(\theta)$ by $\bar{L}_t(\mu) := \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$.

### 3.1. Sequential Variational Approximation

An advantage of variational inference is that it can be directly written as a constrained optimization version of Bayesian inference. To see this we first note that the posterior given in (2) can be obtained by solving the following optimization problem (Dai et al., 2016):

$$p_{t+1}^\eta(\theta) = \operatorname*{arg\,min}_{p\in\mathcal{S}}\left\{\mathbb{E}_{\theta\sim p}\left[\sum_{i=1}^{t}\ell_i(\theta)\right] + \frac{\mathcal{K}(p,\pi)}{\eta}\right\}$$

```
Algorithm 2 Online Variational Inference
Input  Learning rate $\eta > 0$, a prior $\pi(\theta) \in \mathcal{F}$, $q_{\mu_1} \leftarrow \pi$.
For  $t = 1, 2, 3, \dots,$
    1. $\hat{\theta}_t \leftarrow \mathbb{E}_{\theta \sim q_{\mu_t}}[\theta]$,
    2. Observe $\mathcal{D}_t$ to suffer a loss $\ell_t(\hat{\theta}_t)$.
    3. Update depending on the type of algorithm.
        a) For SVA, solve (6).
        b) For SVB, solve (7).
        c) For NGVI, solve (8).
```

We can obtain an approximation by simply restricting the set $\mathcal{S}$:

$$q_{\mu_t} := \operatorname*{arg\,min}_{\mu \in \mathcal{M}} \left\{ \mathbb{E}_{\theta \sim q_\mu} \left[ \sum_{i=1}^{t-1} \ell_i(\theta) \right] + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\} \tag{5}$$

where the set $\mathcal{M}$ is the set of parameters for the set $\mathcal{F} := \{q_\mu, \mu \in \mathcal{M}\}$. The above approximation therefore is a variational approximation of the exact Bayesian inference.

Unfortunately, the update (5) may not be feasible in practice. The Bayesian update of (2) takes a convenient form where update of $p_{t+1}^\eta$ can be written in terms of $p_t^\eta$; see line 3 in Algorithm 1. For update (5), this is not possible in most cases, i.e., we cannot express the optimization problem for $q_{\mu_{t+1}}$ in terms of $q_{\mu_t}$. Typically, one need to store all the past data examples $\mathcal{D}_i$ and recompute their gradients, and then run the optimizer until it converges. This can be very expensive, especially for large $t$.

We propose a sequential version which solves these problems by using an approximation. We follow the ideas used in online gradient algorithms, e.g., such as those used in Shalev-Shwartz (2012), and replace $\mathbb{E}_{\theta \sim q_\mu}[\ell_i(\theta)] = \bar{L}_i(\mu) \approx \mu^T \nabla_\mu \bar{L}_i(\mu_i)$. This leads to

$$\mu_{t+1} = \operatorname*{arg\,min}_{\mu \in \mathcal{M}} \left[ \sum_{i=1}^{t} \mu^T \nabla_\mu \bar{L}_i(\mu_i) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right]. \tag{6}$$

Note that the gradients in the approximation are computed at the past $\mu_i$, rather than the current one $\mu_t$. This results in an algorithm summarized in Algorithm 2 which we call sequential variational approximation (SVA). When computing the gradient of the KL divergence term is feasible, this algorithm can be cheaply performed.

### 3.2. Streaming Variational Bayes

An alternative approach is to remove the term $\mathcal{K}(q_\mu, \pi)$ since $\pi$ is already included in $q_{\mu_t}$:

$$\mu_{t+1} = \operatorname*{arg\,min}_{\mu \in \mathcal{M}} \left[ \mu^T \nabla_\mu \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, q_{\mu_t})}{\eta} \right]. \tag{7}$$

This step, contained in Algorithm 2, is tractable whenever computing the gradient of the KL term is feasible, e.g., when the expectation parameterization is used. This type of update has been proposed in many recent works, e.g., Nguyen et al. (2017a), Zeno et al. (2018). These updates can be seen as a special case of Broderick et al. (2013). Due to this connection, we call this algorithm streaming variational Bayes (SVB).

### 3.3. Natural Gradient Variational Inference

The algorithm described in the previous sections are closely related to existing natural-gradient variational inference (NGVI) algorithm (Sato, 2001; Hoffman et al., 2013; Khan and Lin, 2017). These algorithms are typically applied for *stochastic* learning but can be easily modified for online setting. We will consider the method of Khan and Lin (2017) because it applies to the most general setting (other methods require strong *conjugacy* assumptions on the loss $\ell_t(\theta)$ and prior $\pi(\theta)$). The NGVI algorithm is typically applied to obtain exponential-family approximations, but as we will show the updates are similar to our SVA algorithm which also reveals a more general way of implementing these algorithms in the online setting.

The advantage of using NGVI for online learning is that it obtains closed-form updates for $q_{\mu_{t+1}}$ which can be expressed in terms of $q_{\mu_t}$. This is done by exploiting the expectation parameterization[3] of the exponential family. Throughout this section, we denote the expectation parameter by $\mu$ and natural parameterization of the exponential family by $\lambda$. Khan and Lin (2017) propose the following update[4] in the expectation-parameter space:

$$\min_{\mu \in \mathcal{M}} \left[ \mu^T \nabla_\mu \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} + \frac{\mathcal{K}(q_\mu, q_{\mu_t})}{\alpha} \right], \tag{8}$$

where $\alpha > 0$ is a step size. The difference from (6) is that now the linear term does not contain a sum over all past examples $i$, rather only the current one. Instead, we add another KL divergence term which contains the past information in the previous approximation $q_{\mu_t}$. Therefore, NGVI algorithm, summarized in Algorithm 2, employs a different way to add the past information, but as we show next, it results in a very similar update as SVA. In the appendix, we provide a closed-form solution to (8).

### 3.4. Example: Mean-Field Gaussian VI

We now give a concrete example of the algorithms introduced in this section. We will use the mean-field Gaussian VI where $\mathcal{F}$ is the class of all Gaussian approximations with diagonal covariance matrix. We denote the mean vector of the Gaussian by $m = (m_1, \ldots, m_d)^T$ and the diagonal of the covariance matrix by $\sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2)^T$. To derive the updates for SVA and SVB, we used $\mu = \{m, \sigma\}$ while for NGVI we used the expectation parameters $\mu = \{m, m^2 + \sigma^2\}$. (Here, and until (10) below, the squares and multiplications on vectors

[3] Expectation parameters are expectations of the sufficient statistics, e.g., Gaussian approximation has two expectation parameters: mean vector and correlation matrix respectively.

[4] The exact update proposed in Khan and Lin (2017) is written differently but can be shown to be equivalent to (8). This can be done by using their Lemma 1 and setting $1/\alpha := 1/\beta - 1/\eta$ where $\beta$ is the step-size used in their paper. We use this form since it makes it easier to establish connections to SVA.

are to be understood componentwise). We also assume the prior $\pi(\theta)$ to be a Gaussian with mean 0 and variance $s^2 I_d$ where $I_d$ is the identity $d \times d$ matrix.

Denoting the gradients $\bar{g}_{m_t} := \frac{\partial \bar{L}_t}{\partial m}$ and $\bar{g}_{\sigma_t} := \frac{\partial \bar{L}_t}{\partial \sigma}$, we give the update for each method below (here $h(x) := \sqrt{1+x^2} - x$, applied componentwise for vector inputs):

$$\begin{aligned} \text{SVA: } m_{t+1} &\leftarrow m_t - \eta s^2 \bar{g}_{m_t}, \qquad g_{t+1} \leftarrow g_t + \bar{g}_{\sigma_t}, \\ \sigma_{t+1} &\leftarrow h\left(\tfrac{1}{2}\eta s g_{t+1}\right) s, \end{aligned} \tag{9}$$

$$\begin{aligned} \text{SVB: } m_{t+1} &\leftarrow m_t - \eta \sigma_t^2 \bar{g}_{m_t}, \\ \sigma_{t+1} &\leftarrow \sigma_t h\left(\tfrac{1}{2}\eta \sigma_t \bar{g}_{\sigma_t}\right). \end{aligned} \tag{10}$$

## 4. Generalization Bounds for Online VI

In this section, we present regret bounds for online VI algorithms discussed in the previous section. Our bounds take similar form to the one presented in Theorem 1, and can be used to obtain generazation bounds similar to (4). Our proofs require convexity of $\bar{L}_t(\mu) := \mathbb{E}_{q_\mu}[\ell_t(\theta)]$ with respect to $\mu$, which is a strong assumption. Due to this we are able to derive bounds for SVA and SVB. We expect our bound to hold for NGVI too, due to its similarity to SVA. Specifically, all of our results use the following minimal assumption.

**Assumption 4.1** *$\bar{L}_t$ is L-Lipschitz and convex.*

Some results require the following stronger assumption.

**Assumption 4.2** *$\bar{L}_t$ is $H$-strongly convex where $H > 0$, i.e., for any two $\mu, \mu' \in \mathcal{M}$, the following holds:*

$$\bar{L}_t(\mu') - \bar{L}_t(\mu) \geq (\mu' - \mu)^T \nabla \bar{L}_t(\mu) + \frac{H}{2}\|\mu' - \mu\|^2.$$

Finally, some results also require strong convexity for KL.

**Assumption 4.3** *The KL divergence $\mu \mapsto \mathcal{K}(q_\mu, q_{\mu_1})$ is $\alpha$-strongly convex.*

All of these assumption depend heavily on the parametrization of $\{q_\mu, \mu \in M\}$. For some parameterization, these assumptions do hold although such cases are limited. For example, for Gaussian approximations and convex $\ell$, the assumptions are satisfied, as pointed out by Challis and Barber (2013). This result has recently been extended by Domke (2019) to more generals *location-scale* family. We give a formal statement below.

**Proposition 1 (Theorem 1 in Domke (2019))** *Assuming that $q_\mu$ belongs to a location-scale family $\mathcal{F} = \{q_{m,C}\}$ where $m$ is a $d$-length vector and $C$ is a $d \times d$ matrix with $q_{m,C}(\theta) = [\det(C)]^{-1/2}\psi(C^{-1/2}(\theta - m))$ for some fixed density $\psi$, then $\bar{L}_t$ is convex. Moreover when each $\theta \mapsto \ell_t(\theta)$ is $H$-strongly convex and $\psi$ is the density of a centered random variable with identity variance matrix, then Assumption 4.2 is also satisfied.*

The results for Gaussian approximation can be obtained as a special case.

**Proposition 2** *Assume that $\theta \mapsto \ell_t(\theta)$ is $L'$-Lipschitz. Assume that we use the Gaussian approximation family $\mathcal{F} = \left\{q_{m,C} = \mathcal{N}(m, C^T C), (m, C) \in M\right\}$, $M \subset \mathbb{R}^d \times UT(d)$ where $UT(d)$ is the set of full-rank upper triangular $d \times d$ real matrices. Then $\bar{L}_t$ is $L$-Lipschitz with $L = 2L'$.*

Finally, we remind the formula for the KL divergence between two Gaussian distributions. Let $q_{m,C} = \mathcal{N}(m, C^T C)$ for any $(m, C) \in \mathbb{R}^d \times UT(d)$. Then

$$\mathcal{K}(q_{m,C}, q_{\bar{m},\bar{C}}) = \frac{1}{2}\left( (m - \bar{m})^T \bar{C}^T \bar{C}(m - \bar{m}) + \mathrm{tr}[(\bar{C}^T \bar{C})^{-1}(C^T C)] + \log\left(\frac{\det(\bar{C}^T \bar{C})}{\det(C^T C)}\right) - d \right)$$

is known to be strongly convex on $\mathbb{R}^d \times \mathcal{M}_C$ where $\mathcal{M}_C$ is a closed bounded subset of $UT(d)$. Thus, Assumption 4.3 is satisfied with a Gaussian prior and a Gaussian approximation family.

We are now ready to state our regret bounds for SVA and SVB.

### 4.1. Bounds for SVA

**Theorem 2** *Under Assumptions 4.1 and 4.3, SVA has the following regret bound:*

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) \leq \inf_{\mu \in \mathcal{M}} \left\{ \mathbb{E}_{\theta \sim q_\mu}\left[\sum_{t=1}^{T} \ell_t(\theta)\right] + \frac{\eta L^2 T}{\alpha} + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\}. \tag{11}$$

The above bound is almost identical to the bound given in Theorem 1 where we can replace $p$ by $q_\mu$, $\mathcal{S}$ by $\mathcal{M}$, the bound $B$ by the Lipschitz constant $L$, and factor of 8 by the strong convexity parameter $\alpha$. However, our proof of Theorem 2 is completely different from the one for Theorem 1. It relies on arguments from online convex optimization that can be found in Shalev-Shwartz (2012); Hazan (2016). A detailed proof is given in Appendix 6.5.

Similar to the Bayesian update case discussed in Section 2, using the online-to-batch analysis detailed in Appendix 6.3, we can show that the average $\bar{\theta}_T = (1/T) \sum_{t=1}^{T} \hat{\theta}_t$ satisfies

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{\mu \in \mathcal{M}} \left\{ \mathbb{E}_{\theta \sim q_\mu}[\mathcal{E}_*(\theta)] + \frac{\eta L^2}{\alpha} + \frac{\mathcal{K}(q_\mu, \pi)}{\eta T} \right\}. \tag{12}$$

As an example consider the mean-field Gaussian approximation and assume that for any $\mathcal{D}$, $\ell(\mathcal{D}, \cdot)$ is $L/2$-Lipschitz (note that these are the assumptions of Proposition 2 ensuring that Assumption 4.1 is satisfied). Then $\mathbb{E}_{\theta \sim q_\mu}[\mathcal{E}_*(\theta)] = \mathcal{E}_*(m) + \|\sigma\| L/2$ . Therefore, given the expression of the KL-divergence between Gaussian distributions, taking a vector $\sigma$ with $\sigma_j = L\eta/(\alpha\sqrt{d})$, $\eta = (1/L)\sqrt{\alpha d \log(T/d)/T}$, and considering only the regret with respect to bounded means $m$ leads to

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{m \in [-\bar{M}, \bar{M}]^d} \mathcal{E}_*(m) + (1 + o(1)) \frac{2L}{\alpha} \sqrt{\frac{d \log(dT)}{T}}.$$

This again is very similar to the generalization error shown in (4).

### 4.2. Bounds for SVB

Similarly to the SVA case, we can derive a regret bound, however our proof only applies to the Gaussian case. For this case, we require a dynamic learning $\eta_t$. We use a different learning rate for each element of $\theta_j$ which we denote by $\eta_{t,j}$. The result also works for a bounded parameter space $\mathcal{M} = \mathcal{M}_m \times \mathcal{M}_\sigma$ that will imply a projection step in addition to the update in (10):

$$\begin{aligned} \text{SVB: } m_{t+1} &\leftarrow \Pi_{\mathcal{M}_m}\left[m_t - \eta\sigma_t^2 \bar{g}_{m_t}\right], \\ \sigma_{t+1} &\leftarrow \Pi_{\mathcal{M}_\sigma}\left[\sigma_t h\left(\tfrac{1}{2}\eta\sigma_t \bar{g}_{\sigma_t}\right)\right]. \end{aligned}$$

where $\Pi_{\mathcal{M}_m}$ and $\Pi_{\mathcal{M}_\sigma}$ denote the orthogonal projection on $\mathcal{M}_m$ and $\mathcal{M}_\sigma$ respectively. The following theorem states the result.

**Theorem 3** *We consider the mean-field Gaussian family $q_\mu = \mathcal{N}(m, \mathrm{diag}(\sigma^2))$ and $\mathcal{M} = \mathcal{M}_m \times \mathcal{M}_\sigma$ where $\mathcal{M}_m$ and $\mathcal{M}_\sigma$ are closed, bounded, convex subsets of $\mathbb{R}^d$ and $\mathbb{R}_+^d$ respectively, and $0 \in \mathcal{M}_\sigma$. Define $D^2 = \sup\left\{\|m - m'\|_2^2 + \|\sigma\|^2, m, m' \in \mathcal{M}_m, \sigma \in \mathcal{M}_\sigma\right\}$. Then, under Assumption 4.1, with the choice $\eta_{t,j} = \frac{D\sqrt{2}}{L}\frac{1}{\sqrt{t}\sigma_{t,j}^2}$ we get:*

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) \leq \inf_{\theta \in \mathcal{M}_m} \sum_{t=1}^{T} \ell_t(\theta) + DL\sqrt{2T}. \tag{13}$$

*Under Assumptions 4.1 and 4.2, the choice $\eta_t = 2/Ht\sigma_t^2$ leads to:*

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) \leq \inf_{\theta \in \mathcal{M}_m} \sum_{t=1}^{T} \ell_t(\theta) + \frac{L^2(1 + \log T)}{H}. \tag{14}$$

Here again the results are similar to the Bayesian inference case but now expressed in terms of the parameters $\mu$ instead of expectations.

A similar bound on the generalization error can also be proved. Define $\bar{\theta}_T = (1/T)\sum_{t=1}^{T} \hat{\theta}_t$. Here, the online-to-batch analysis directly leads to

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{\theta \in \mathcal{M}_m} \mathcal{E}_*(\theta) + \frac{DL\sqrt{2}}{\sqrt{T}}$$

in the convex case and

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{\theta \in \mathcal{M}_m} \mathcal{E}_*(\theta) + \frac{L^2(1 + \log T)}{HT}$$

in the strongly convex case.

Note the in the online optimization setting studied in Shalev-Shwartz (2012), it is usual to optimize on Euclidean balls. Here, $M_m = \{m \in \mathbb{R}^d : \|m\| \leq \bar{M}\}$ and $M_\sigma = \{\sigma \in \mathbb{R}_+^d : \|\sigma\| \leq \bar{S}\}$ leads to $D = 4\bar{M}^2 + \bar{S}^2$ leads to dimension-free bounds.

On the other hand, the choice $M_m = [-\bar{M}, \bar{M}]^d$ and $M_\sigma = [0, \bar{S}]^d$ implies $D^2 = d(4\bar{M}^2 + \bar{S}^2)$, and so the bound in the convex case is

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{\theta \in \mathcal{M}_m} \mathcal{E}_*(\theta) + \frac{L\sqrt{2d(4\bar{M}^2 + \bar{S}^2)}}{\sqrt{T}}$$

and its dependence in $d$ is the same as in the bound on SVA.

### 4.3. Generalization

We expect our bounds to hold for NGVI as well. When expectation parameterization is used, the assumptions are satisfied only in very limited models. This is because the result of Proposition 1 and 2 do not directly apply to expectation parameterization. However, the NGVI update shown in (8) can be applied in other parameterization as well, in which case some of our result can be extended to NGVI too.

## 5. Experiments

In this section, we conduct experiments on real and simulated datasets, in classification and linear/nonlinear regression. The objective is twofold: check the convergence of SVA/SVB, with and without the convexity assumption on $\bar{L}_t$, and compare SVA, NGVI and SVB.

### 5.1. Experimental setup

We compare the empirical performance of the algorithms we present in this paper through classification and regression tasks on several toy and real-world datasets. We also include the classical online gradient descent and the online gradient descent on the expected loss as benchmarks. Please refer to Appendix 6.2 for more details on these algorithms. In the following, OGA will stand for the classical online gradient descent while OGA-EL for the OGA on the expected loss (Algorithm 3). We recall that SVA, NGVI and SVB respectively refer to the sequential variational approximation (6), natural gradient variational inference (8) and streaming variational Bayes (7).

**Binary classification** We consider first a classification problem. At each round $t$ the learner receives a data point $x_t \in \mathbb{R}^d$ and predicts its label $y_t \in \{-1, +1\}$ using $\langle x_t, \theta_t \rangle$. The adversary reveals the true value $y_t$, then the learner suffers the loss $\ell_t(\theta_t) = (1 - y_t \theta_t^T x_t)_+$, where $a_+ = a$ if $a > 0$ and $a_+ = 0$ otherwise.

**Regression** At each round $t$, the learner receives a set of features $x_t \in \mathbb{R}^d$ and predicts $y_t \in \mathbb{R}$ using $\langle x_t, \theta_t \rangle$. Then the adversary reveals the true value $y_t$ and the learner suffers the loss $\ell_t(\theta_t) = (y_t - f_{\theta_t}(x_t))^2$. We will consider both the linear case when the predictions are linear $f_\theta(x_t) = \theta^T x_t$ and the nonlinear case where the predictions are outputs of a one-hidden-layer neural network with a ReLU activation. The first case of linear predictions leads to a convex loss with respect to $\theta$, while the latter leads to a nonconvex loss.

**Variational family** For both tasks, we use a Gaussian mean-field variational family $\mathcal{F} = \{q_\mu = \mathcal{N}\left(m, \operatorname{diag}(\sigma^2)\right) / \mu = (m, \sigma) \in M_m \times M_\sigma\}$, $M_m = [-20, 20]^d$ and $M_\sigma = [0, 1]^d$.

**Datasets** We consider here six different datasets: one toy and three real datasets for classification, and one real world dataset for both linear and nonlinear regression. The three real world datasets used for the binary classification problem are the popular Breast Cancer, the Pima Indians and the Forest Cover Type datasets, while those used for regression are the Boston Housing and the California Housing datasets respectively for the convex and the nonconvex case. All come from the UCI machine learning repository. Note that in some databases, the data are ordered according to some criterion such as the date or the label. In order to avoid any effect linked to this, we randomly permuted the observations.

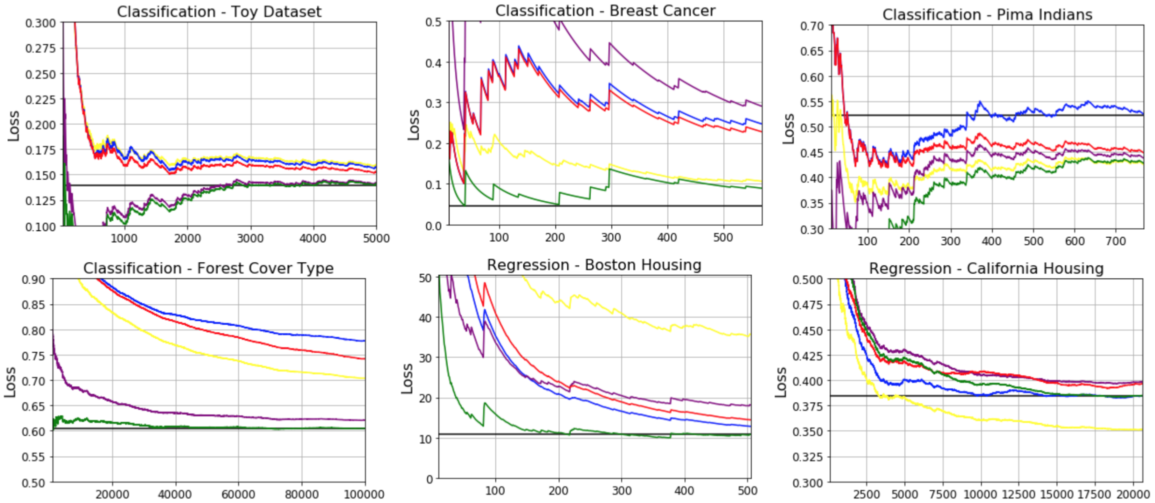

Figure 1: Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green) for the convex hinge loss and the squared loss functions. The black line shows the average total cumulative loss in hindsight. We see that in most cases NGVI outperforms the other algorithms. The last plot (California Housing dataset) shows the consistency of our algorithms for a nonconvex loss $\bar{L}_t$.

The toy dataset is as follows: we sample $n = 10^4$ points $y_t$ according to a Bernoulli distribution $\mathcal{B}e(2/3)$. Then

$$x_t|(y_t = 1) \sim \mathcal{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}\right) \text{ and } x_t|(y_t = 0) \sim \mathcal{N}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

| Dataset | $T$ | $d$ | Dataset | $T$ | $d$ |
|---|---|---|---|---|---|
| Toy classification | 10000 | 2 | Cover Type | 581012 | 54 |
| Breast cancer | 569 | 30 | Boston Housing | 506 | 13 |
| Pima Indians | 768 | 8 | California Housing | 20640 | 9 |

### 5.2. Experimental results

For each task and each dataset, we plot the evolution of the average cumulative loss $\sum_{i=1}^{t} \ell_i(\theta_i)/t$ as a function of the step $t = 1, ..., T$, where $T$ is the number of instances of the dataset and $\theta_i$ is the decision made by the learner at step $i$. We compare this quantity to the best average total cumulative loss in hindsight $\inf_{\theta \in M_m} \frac{1}{T} \sum_{t=1}^{T} \ell_t(\theta)$ which is represented by a straight black horizontal line in Figure 1.

**Parameters setting** We initialize all means to 0 and all values of the variance to 1. For simplicity, the values of the learning rates are set to $\eta = 1/\sqrt{T}$ for OGA, OGA-EL and SVA while $\eta_t = 1/\sigma_t^2\sqrt{t}$ for SVB and $\eta_t = 1$ for NGVI respectively. It is possible to optimize the values of the step sizes. Nevertheless, we draw attention to the fact that a simple cross validation technique would not be valid here as it would require to know the whole dataset before selecting the step size, which is not possible in an online setting, and using such a strategy at each step $t$ using the past data would change the learning rate of OGA, OGA-EL and SVA at each step.

**Conclusions** The results are reported in Figure 1 that shows the consistency of our algorithms. The goal of our simulations is to observe the empirical performance of our algorithms in practice, and to see if it is possible to go further than the convexity assumption that is required in Section 4. Looking at the plots, the two main findings of our experiments are the following:

- the generalization properties of online variational inference seem to go beyond the convex assumption we stated in the previous theoretical parts.
- even though SVA and SVB exhibit good performances, NGVI is the best method in practice as it converges faster on all the datasets.

## 6. Conclusion

In this paper, we derive the first generalization bounds for some online variational inference algorithms. Our proof techniques applies to cases where existing methods do not work. By using existing variational methods, we proposed a few online methods for variational inference. We provided generalization bounds for the SVA algorithm, and related them to the NGVI methods. We also derived a bound for a special case of SVB. We provided numerical results to establish consistency of our results. We observed that NGVI outperforms all the other methods, and that the theoretical convexity assumption is not needed in practice.

We believe that it is possible to extend our proof techniques to NGVI case. Currently, our proofs strongly rely on the convexity of $\mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ with respect to $\mu$. This analysis cannot directly be used for the parameterization of Khan and Lin (2017). However, it can be applied to a general formulation where our assumptions hold. We believe that generalization bounds for NGVI is possible to derive and will pursue this direction in the future.

## References

P. Alquier and J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics (to appear)*, 2017.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *JMLR*, 17(239):1–41, 2016.

J. Y. Audibert. Fast learning rates in statistical inference through aggregation. *Annals of Statistics*, 37(4):1591–1646, 2009.

A. Banerjee. On Bayesian bounds. In *Proceedings of ICML*, pages 81–88. ACM, 2006.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Statistics*, 37(6):1554–1563, 12 1966.

A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *arXiv preprint arXiv:1611.01125, to appear in the Annals of Statistics*, 2016.

A. Bhattacharya, D. Pati, and Y. Yang. On statistical optimality of variational Bayes. *PMLR: Proceedings of AISTAT*, 84, 2018.

D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM, 2006.

T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational Bayes. In *NIPS*, pages 1727–1735. Curran Associates, Inc., 2013.

S. Bubeck. Introduction to online optimization. Lecture notes (Princeton University), 2011.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Saint-Flour Summer School on Probability Theory 2001, Lecture Notes in Mathematics. Springer, 2004.

O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. IMS Lecture Notes, Monograph Series, 56. 2007.

O. Catoni and I. Giulini. Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector. NIPS Workshop: Almost 50 Shades of Bayesian Learning, 2017.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

E. Challis and D. Barber. Gaussian Kullback-Leibler approximate inference. *JMLR*, 14(1): 2239–2286, 2013.

B.-E. Chérief-Abdellatif. Consistency of ELBO maximization for model selection. *PMLR: Proceedings of AABI*, 96:11–31, 2019.

B.-E. Chérief-Abdellatif and P. Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12(2):2995–3035, 2018.

V. Cottet and P. Alquier. 1-bit matrix completion: PAC-Bayesian analysis of a variational approximation. *Machine Learning*, 107(3):579–603, 2018.

N. V. Cuong, L. S. T. Ho, and V. Dinh. Generalization and robustness of batched weighted average algorithm with V-geometrically ergodic Markov data. In *International Conference on Algorithmic Learning Theory*, pages 264–278. Springer, 2013.

B. Dai, N. He, H. Dai, and L. Song. Provable Bayesian inference via particle mirror descent. In *AISTAT*, pages 985–994, 2016.

J. Domke. Provable smoothness guarantees for black-box variational inference. preprint ArXiv., 2019.

A. Doucet and A. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12, 01 2009.

S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *JMLR*, 14(1):729–769, 2013.

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

S. Ghosal and A. Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

P. D. Grünwald and T. Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

B. Guedj. A primer on PAC-Bayesian learning. *Preprint arXiv:1901.05353*, 2019.

R. Guhaniyogi, R. M. Willett, and D. B. Dunson. Approximated Bayesian inference for massive streaming data. Unpublished manuscript, 2013.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3–4):157–325, 2016.

M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

P. Jaiswal, V. A. Rao, and H. Honnappa. Asymptotic consistency of $\alpha$-rényi-approximate posteriors. Preprint arXiv:1902.01902, 2019.

M. I. Jordan, Z. Ghahramani, Tommi S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

M. E. Khan and W. Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *PMLR: Proceedings of ICML*, 54:878–887, 2017.

M. E. Khan and D. Nielson. Fast yet simple natural-gradient descent for variational inference in complex models. Invited paper at ISITA 2018, 2018.

M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in ADAM. ICML, 2018.

Di. Kingma and M. Welling. Auto-encoding variational Bayes. *Preprint arXiv:1312.6114*, 2013.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

C. V. Nguyen, T. D. Bui, Y. Li, and R. E. Turner. Online variational Bayesian inference: Algorithms for sparse Gaussian processes and theoretical bounds. ICML 2017 Time Series Workshop, 2017a.

C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. Variational continual learning. *Preprint arXiv:1710.10628*, 2017b.

J. Rousseau. On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3:211–231, 2016.

M.-A. Sato. Online model selection based on the variational Bayes. *Neural computation*, 13 (7):1649–1681, 2001.

Y. Seldin and N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *JMLR*, 11: 3595–3646, 2010.

Y. Seldin, P. Auer, J. Shawe-Taylor, R. Ortner, and F. Laviolette. Pac-bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems*, pages 1683–1691, 2011.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayesian estimator. In *Tenth annual conference on Computational learning theory*, volume 6, pages 2–9, 1997.

R. Sheth and R. Khardon. Excess risk bounds for the Bayes risk using variational inference in latent Gaussian models. In *NIPS*, pages 5151–5161, 2017.

T. Suzuki. PAC-Bayesian bound for Gaussian process regression and multiple kernel additive model. In *Conference on Learning Theory*, pages 8–1, 2012.

Y. Tsuzuku, I. Sato, and M. Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. *Preprint arXiv:1901.04653*, 2019.

V. G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, 1990.

C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. In *Proceedings of AISTAT 2011*, pages 752–760, 2011.

Y. Wang and D. M. Blei. Frequentist consistency of variational Bayes. Journal of the American Statistical Association (to appear), 2018.

C. Zeno, I. Golan, E. Hoffer, and D. Soudry. Bayesian gradient descent: Online variational Bayes learning with increased robustness to catastrophic forgetting and weight pruning. Preprint arXiv:1803.10123, 2018.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *Preprint arXiv:1712.02519v1*, 2017.

## Appendix

### 6.1. Closed-form solutions for NGVI

The expectation parameterization of NGVI enables closed-form solution. This is because the gradient of the KL diverence with respect to expectation parameter is available in closed-form (see Eq. 10 in Khan and Nielson (2018)). The closed update is given in Eq. (50) in Khan and Lin (2017) using which we obtain the following update:

$$\lambda_{t+1} = (1-\beta)\lambda_t + \beta\lambda_1 - \eta\beta\nabla_\mu \bar{L}_t(\mu_t), \tag{15}$$

where $1/\beta := 1/\alpha + 1/\eta$. Given $\lambda_{t+1}$, we can get $\mu_{t+1} = \nabla_\lambda A(\lambda_{t+1})$ where $A$ is the log-partition function of the exponential family.

Now we show that this closed-form update is similar to SVA. By using induction similar to Lemma 4 in Khan and Lin (2017), we can write the update in terms of all past gradients:

$$\lambda_{t+1} = \lambda_1 - \eta \sum_{i=1}^{t} w_i \nabla_\mu \bar{L}_i(\mu_i) \tag{16}$$

where $w_i := \beta \prod (1-\beta)^{i-2}$. This can be compared to the SVA update in the expectation parameterization where applying the gradient to (6) gives us the following update similar to (15) but where $w_i = 1$ for all $i$:

$$\lambda_{t+1} = \lambda_1 - \eta \sum_{i=1}^{t} \nabla_\mu \bar{L}_i(\mu_i) \tag{17}$$

Therefore, SVA takes a gradient step assuming that all gradients are equally important, which is similar to the Bayesian update (2) where all loss $\ell_i$ are treated equally. In contrast, in NGVI, the past gradients are discounted using $\beta$ and ultimately forgotten. Weighting past gradients makes sense when we do not want the current mistakes to affect the future. However, the choice of step-size is crucial to know the rate at which the past gradients should be discounted.

NGVI is typically applied using expectation parameterization, but the formulation (8) is more general although could be computationally difficult. The theoretical results in the paper further assume that $\bar{L}_i$ is convex in $\mu$. Still, in our experiments, NGVI gives good performance in an online setting compared to many other algorithms.

### 6.2. Online gradient algorithm on the expected loss (OGA-EL)

It is possible to directly use the online gradient algorithm (OGA) on the expected loss $\mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$, see Algorithm 3.

Note first that from Shalev-Shwartz (2012) step (iii) is actually equivalent to

$$\mu_{t+1} = \arg\min_{\mu\in M} \left[ \sum_{i=1}^{t} \mu^T \nabla \bar{L}_i(\mu_i) + \frac{\|\mu-\mu_1\|^2}{\eta} \right],$$

which means that we replaced the Küllback-Leibler divergence by the Euclidean norm in SVA.

```
Algorithm 3 OGA-EL
Input  Learning rate $\eta > 0$, a prior $\pi(\theta) \in \mathcal{F}$, $q_{\mu_1} \leftarrow \pi$.
Loop  For $t = 1, \dots,$
    1. $\hat{\theta}_t \leftarrow \mathbb{E}_{\theta \sim q_{\mu_t}}[\theta]$,
    2. Observe $\mathcal{D}_t$ to suffer a loss $\ell_t(\hat{\theta}_t)$.
    3. Update $\mu_{t+1} = \mu_t - \eta \nabla \bar{L}_t(\mu_t)$.
```

Also, when $\mu = (m, \sigma) \in \mathbb{R}^d \times (\mathbb{R}_+)^d$ and $q_\mu = \mathcal{N}(m, \text{diag}(\sigma))$, then Algorithm 3 becomes

$$m_{t+1} = m_t - \eta s^2 \frac{\partial \bar{L}_t}{\partial m}(m_t, \sigma_t),$$
$$\sigma_{t+1} = \sigma_t - \eta s^2 \frac{\partial \bar{L}_t}{\partial \sigma}(m_t, \sigma_t).$$

We have regret bounds for this method, similar to the one for EWA:

**Theorem 4** *Under Assumption 4.1, Algorithm 3 leads to:*

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) \leq \inf_{\mu \in M} \left\{ \mathbb{E}_{\theta \sim q_\mu} \left[ \sum_{t=1}^{T} \ell_t(\theta) \right] + \eta L^2 T + \frac{\|\mu - \mu_1\|^2}{\eta} \right\},$$

*and moreover, under Assumptions 4.3 and 4.1, Algorithm 3 leads to:*

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) \leq \inf_{\mu \in M} \left\{ \mathbb{E}_{\theta \sim q_\mu} \left[ \sum_{t=1}^{T} \ell_t(\theta) \right] + \eta L^2 T + \frac{\alpha \mathcal{K}(q_\mu, \pi)}{2\eta} \right\}.$$

The proof of this result is given below with the other proofs of the paper.

### 6.3. Online-to-batch conversion

Many times in the paper, we derived generalization error bounds from regret bounds, using the online-to-batch conversion. We here give a formal statement for this result, note that this result is essentially Theorem 5.1 in Shalev-Shwartz (2012). We also provide a proof for the sake of completeness.

**Theorem 5** *Assume that $\mathcal{D}_1, \dots, \mathcal{D}_T$ are i.i.d from $P_*$. Assume we use an online algorithm on the data that produce a sequence of parameters $\hat{\theta}_1, \dots, \hat{\theta}_T$. That is, $\hat{\theta}_t = \hat{\theta}(\mathcal{D}_1, \dots, \mathcal{D}_{t-1})$. Define the estimator*

$$\bar{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \hat{\theta}_t.$$

*Then*

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \mathbb{E}_{\mathcal{D}_{1:T} \sim P_*} \left[ \frac{1}{T} \sum_{t=1}^{T} \ell_t(\hat{\theta}_t) \right].$$

**Proof** We have:

$$\mathcal{E}_*(\bar{\theta}_T) = \mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell(\mathcal{D},\bar{\theta}_T)\right] = \mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell\left(\mathcal{D},\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t\right)\right] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell\left(\mathcal{D},\hat{\theta}_t\right)\right]$$

by Jensen's inequality. The key is that as $\hat{\theta}_t = \hat{\theta}_t(\mathcal{D}_1,\dots,\mathcal{D}_{t-1})$ does not depend on $\mathcal{D}_t$, we can rewrite:

$$\mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell\left(\mathcal{D},\hat{\theta}_t\right)\right] = \mathbb{E}_{\mathcal{D}_t\sim P_*}\left[\ell\left(\mathcal{D}_t,\hat{\theta}_t\right)\right] = \mathbb{E}_{\mathcal{D}_t\sim P_*}\left[\ell_t(\hat{\theta}_t)\right]$$

and so we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\mathcal{E}_*(\bar{\theta}_T)\right] &\leq \mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell_t(\hat{\theta}_t)\right]\right\} \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\ell_t(\hat{\theta}_t)\right] = \mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_t(\hat{\theta}_t)\right].\end{aligned}$$

■

As an application, we state an exact version of (4) and prove it from Theorem 1 and Theorem 5.

**Theorem 6** *Assume that the loss $\ell$ is bounded by $B$ as in Theorem 1 and that $\mathcal{D}_1,\dots,\mathcal{D}_T$ are i.i.d from $P_*$. Assume that there is some $d>0$ such that*

$$r(\varepsilon) \leq -d\log(1/\varepsilon)$$

*where $r(\varepsilon) = \log[1/\pi(B(\theta^*,\varepsilon))]$ and $B(\theta^*,\varepsilon) = \{\theta\in\Theta : \mathcal{E}(\theta)-\mathcal{E}(\theta^*)\leq\varepsilon\}$. Use on this data the EWA strategy with $\eta = (1/2\sqrt{2}B)\sqrt{(d/T)\log(d/T)}$, then*

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\hat{\theta}_T)] \leq \mathcal{E}_*(\theta^*) + B\sqrt{\frac{d}{2T}\log\left(\frac{T}{d}\right)} + \frac{d}{T}.$$

Note that the prior mass condition is classical in the PAC-Bayesian literature and in the frequentist analysis of Bayesian estimators, see e.g Catoni (2007); Rousseau (2016); Bhattacharya et al. (2016); Ghosal and Van der Vaart (2017). The estimator $\bar{\theta}_T$ averaging the decisions $\hat{\theta}_t$ was first introduced by Catoni (2004) as the "double mixture rule".

**Proof** Define $p_\varepsilon$ as $\pi$ restricted to $B(\theta^*,\varepsilon)$ and note that

$$\mathcal{K}(p_\varepsilon,\pi) = -\log B(\theta^*,\varepsilon) = r(\varepsilon) \leq d\log(1/\varepsilon).$$

From Theorem 1, for any $\varepsilon$,

$$\sum_{t=1}^{T}\ell_t(\hat{\theta}_t) \leq \mathbb{E}_{\theta\sim p_\varepsilon}\left[\sum_{t=1}^{T}\ell_t(\theta)\right] + \frac{\eta B^2 T}{8} + \frac{d\log(1/\varepsilon)}{\eta}.$$

From Theorem 5,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\hat{\theta}_T)] &= \mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_t(\hat{\theta}_t)\right] \\
&\leq \mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left\{\mathbb{E}_{\theta\sim p_\varepsilon}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_t(\theta)\right]\right\} + \frac{\eta B^2}{8} + \frac{d\log(1/\varepsilon)}{T\eta} \\
&= \mathbb{E}_{\theta\sim p_\varepsilon}\left[\mathcal{E}_*(\theta)\right] + \frac{\eta B^2}{8} + \frac{d\log(1/\varepsilon)}{T\eta} \\
&\leq \mathcal{E}_*(\theta^*) + \varepsilon + \frac{\eta B^2}{8} + \frac{d\log(1/\varepsilon)}{T\eta}
\end{aligned}$$

where the last inequality comes from the definition of $p_\varepsilon$. Taking $\varepsilon = d/T$ gives:

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\hat{\theta}_T)] \leq \mathcal{E}_*(\theta^*) + \frac{d}{T} + \frac{\eta B^2}{8} + \frac{d\log(T/d)}{T\eta}.$$

Finally, substitute its value to $\eta$ to get

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\hat{\theta}_T)] \leq \mathcal{E}_*(\theta^*) + B\sqrt{\frac{d}{2T}\log\left(\frac{T}{d}\right)} + \frac{d}{T}.$$

■

### 6.4. A tool for the proofs

We remind the following classical lemma. We refer the reader for example to Catoni (2007) for a proof of this result, where it is stated as Lemma 1.1.3 (page 16).

**Lemma 1** *Let $h : \Theta \to \mathbb{R}$ be a bounded measurable function and $\pi \in \mathcal{S}(\Theta)$. Then*

$$\sup_{p\in\mathcal{S}(\Theta)} \left\{\mathbb{E}_{\theta\sim p}[h(\theta)] - \mathcal{K}(p,\pi)\right\} = \log \mathbb{E}_{\theta\sim\pi}[\exp(h(\theta))]$$

*and the supremum is actually reached for*

$$p(\theta) \propto \exp[h(\theta)]\pi(\theta).$$

This lemma will actually turn out to be a fundamental tool for some of the proofs.

### 6.5. Proofs

**Proof** [Proof of Theorem 1] Note that this proof is classical and is reminded here for the sake of completeness. We have first:

$$\begin{aligned}
\exp\left[-\eta\ell_t(\hat{\theta}_t)\right] &= \exp\left[-\eta\ell_t(\mathbb{E}_{\theta\sim p_t^\eta}(\theta))\right] \\
&\geq \exp\left[-\eta\mathbb{E}_{\theta\sim p_t^\eta}(\ell_t(\theta))\right]
\end{aligned}$$

$$\geq \mathbb{E}_{\theta\sim p_t^\eta}\left\{\exp\left[-\eta\ell_t(\theta)-\frac{\eta^2 B^2}{8}\right]\right\}$$

where we used respectively Jensen and Hoeffding's inequality. So

$$\ell_t(\hat{\theta}_t) \leq \frac{\eta B^2}{8} - \frac{1}{\eta}\log \mathbb{E}_{\theta\sim p_t^\eta}\exp\left[-\eta\ell_t(\theta)\right]. \tag{18}$$

Remind that by definition,

$$p_t^\eta(\theta) = \frac{\exp\left(-\eta\sum_{i=1}^{t-1}\ell_i(\theta)\right)\pi(\theta)}{N_t}$$

where $N_t$ is the normalisation constant given by

$$N_t = \mathbb{E}_{\theta\sim\pi}\left[\exp\left(-\eta\sum_{i=1}^{t-1}\ell_i(\theta)\right)\right].$$

But note that then

$$\log \mathbb{E}_{\theta\sim p_t^\eta}\exp\left[-\eta\ell_t(\theta)\right] = \log\left(\frac{N_{t+1}}{N_t}\right).$$

We plug this into (18) and sum for $t = 1,\dots,T$. We obtain

$$\begin{aligned}\sum_{t=1}^T \ell_t(\hat{\theta}_t) &\leq \frac{\eta B^2 T}{8} - \frac{1}{\eta}\sum_{t=1}^T \log\left(\frac{N_{t+1}}{N_t}\right)\\ &= \frac{\eta B^2 T}{8} - \frac{1}{\eta}\log\left(\frac{N_{T+1}}{N_1}\right)\\ &= \frac{\eta B^2 T}{8} - \frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\pi}\left[\exp\left(-\eta\sum_{t=1}^T\ell_t(\theta)\right)\right]\right).\end{aligned}$$

Lemma 1 leads to

$$\sum_{t=1}^T \ell_t(\hat{\theta}_t) \leq \frac{\eta B^2 T}{8} + \inf_{p\in\mathcal{S}(\Theta)}\left\{\mathbb{E}_{\theta\sim p}\left[\sum_{t=1}^T\ell_t(\theta)\right] + \frac{\mathcal{K}(p,\pi)}{\eta}\right\}.$$

■

**Proof** [Proof of Proposition 2] Let $\varphi_{m,C}(\cdot)$ denote the p.d.f of the Gaussian distribution with mean $m$ and variance matrix $C$. Let $(m_1,C_1),(m_2,C_2)\in M$,

$$\begin{aligned}|\bar{L}_t(m_1,C_1)-\bar{L}_t(m_2,C_2)| &= \left|\int \ell_t(\theta)\varphi_{m_1,C_1}(\theta)\mathrm{d}\theta - \int \ell_t(\theta)\varphi_{m_2,C_2}(\theta)\mathrm{d}\theta\right|\\ &\leq \int \left|\ell_t(m_1+C_1u)-\ell_t(m_2+C_2u)\right|\varphi_{0,I_d}(u)\mathrm{d}u\\ &\leq L'\|m_1-m_2\| + L'\int \|(C_1-C_2)u\|\varphi_{0,I_d}(u)\mathrm{d}u.\end{aligned}$$

For any $C = (C_{i,j}) \in UT(d)$, we have

$$\int \|Cu\| \varphi_{0,I_d}(u)\mathrm{d}u \leq \sqrt{\int \|Cu\|^2 \varphi_{0,I_d}(u)\mathrm{d}u}$$
$$= \sqrt{\int \sum_{i=1}^{d} \left(\sum_{j=1}^{d} C_{i,j} u_j\right)^2 \varphi_{0,I_d}(u)\mathrm{d}u} = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} C_{i,j}^2}$$

which leads to

$$|\bar{L}_t(m_1, C_1) - \bar{L}_t(m_2, C_2)| \leq L'\|m_1 - m_2\| + L' \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} (C_1 - C_2)_{i,j}^2}$$
$$\leq 2L'\|(m_1, C_1) - (m_2, C_2)\|.$$

This ends the proof. ■

**Proof** [Proof of Theorem 2] First, Assumption 4.1 ensures that the $\bar{L}_t$'s are convex. By definition of the subgradient of a convex function,

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)] &= \sum_{t=1}^{T} \ell_t\left(\mathbb{E}_{\theta\sim q_{\mu_t}}(\theta)\right) - \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)] \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu_t}}[\ell_t(\theta)] - \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)] \\
&= \sum_{t=1}^{T} \bar{L}_t(\mu_t) - \sum_{t=1}^{t} \bar{L}_t(\mu) \\
&\leq \sum_{t=1}^{T} \mu_t^T \nabla \bar{L}_t(\mu_t) - \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t). \qquad (19)
\end{aligned}$$

Then, following the general proof scheme detailed in Chapter 2 in Shalev-Shwartz (2012), we prove by recursion on $T$ that for any $\mu \in \mathcal{M}$,

$$\sum_{t=1}^{T} \mu_t^T \nabla \bar{L}_t(\mu_t) - \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t) \leq \sum_{t=1}^{T} \mu_t^T \nabla \bar{L}_t(\mu_t) - \sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \qquad (20)$$

which is exactly equivalent to

$$\sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \leq \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta}. \qquad (21)$$

Indeed, for $T = 0$, (21) just states that $\mathcal{K}(q_\mu, \pi) \geq 0$ which is a well-known property of KL. Assume that (21) holds for some integer $T - 1$. We then have, for all $\mu \in M$,

$$\sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) = \sum_{t=1}^{T-1} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) + \mu_{T+1}^T \nabla \bar{L}_T(\mu_T)$$

$$\leq \sum_{t=1}^{T-1} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} + \mu_{T+1}^T \nabla \bar{L}_T(\mu_T)$$

as (21) holds for $T-1$. Apply this to $\mu = \mu_{T+1}$ to get

$$\begin{aligned}
\sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) &\leq \sum_{t=1}^{T} \mu_{T+1}^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(p_{\mu_{T+1}}, \pi)}{\eta} \\
&= \min_{m \in \mathcal{M}} \left[ \sum_{t=1}^{T} m^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(p_m, \pi)}{\eta} \right], \text{ by definition of } \mu_{T+1} \\
&\leq \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta}
\end{aligned}$$

for all $\mu \in \mathcal{M}$. Thus, (21) holds for $T$. Thus, by recursion, (21) and (20) hold for all $T \in \mathbb{N}$.

The last step is to prove that for any $t \in \mathbb{N}$,

$$\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \leq \frac{\eta L^2}{\alpha}. \tag{22}$$

Indeed,

$$\begin{aligned}
\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) &= (\mu_t - \mu_{t+1})^T \nabla \bar{L}_t(\mu_t) \\
&\leq \|\mu_t - \mu_{t+1}\| \|\nabla \bar{L}_t(\mu_t)\| \text{ by Cauchy-Schwarz} \\
&\leq L \|\mu_t - \mu_{t+1}\| \qquad (23)
\end{aligned}$$

as $\bar{L}_t$ is $L$ Lipschitz (Assumption 4.1). Define

$$G_t(\mu) = \sum_{i=1}^{t-1} \mu^T \nabla \bar{L}_i(\mu_i) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta}.$$

Note that from Assumption 4.3, $\mu \mapsto \mathcal{K}(q_\mu, \pi)/\eta$ is $\alpha/\eta$-strongly convex. As the sum of a linear function and an $\alpha/\eta$-strongly convex function, $G_t$ is $\alpha/\eta$-strongly convex. So, for any $(\mu, \mu')$,

$$G_t(\mu') - G_t(\mu) \geq (\mu' - \mu)^T \nabla G_t(\mu) + \frac{\alpha \|\mu' - \mu\|^2}{2\eta}.$$

As a special case, using the fact that $\mu_t$ is a minimizer of $G_t$, we have

$$G_t(\mu_{t+1}) - G_t(\mu_t) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

In the same way,

$$G_{t+1}(\mu_t) - G_{t+1}(\mu_{t+1}) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

Summing the two previous inequalities gives

$$\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{\eta},$$

and so, combined with, this gives:

$$\|\mu_{t+1} - \mu_t\| \leq \sqrt{\frac{\eta}{\alpha}\left[\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t)\right]}.$$

Combining this inequality with (23) leads to (22).

Plugging (19), (20) and (22) together gives

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] \leq \frac{\eta T L^2}{\alpha} + \frac{\mathcal{K}(q_\mu, \pi)}{\eta},$$

that is the statement of the theorem. ■

**Proof** [Proof of Theorem 3] We prove this theorem from scratch and use the main techniques outlined in Hazan (2016). As previously, the idea is to study differences $\bar{L}_t(\mu_t) - \bar{L}_t(\mu)$. However, in this case, we have, for any $\mu = (m, \sigma)$, using Jensen's inequality,

$$\bar{L}_t(m, \sigma) = \mathbb{E}_{\theta \sim q_{m,\sigma}}[\ell_t(\theta)] \geq \ell_t(m) = \bar{L}_t(m, 0).$$

So, we can assume from the beginning that $\mu = (m, 0)$.

**Convex case:**

First, we assume that each function $\bar{L}_t$ is convex, for all $m = (m_1, ..., m_d) \in \mathcal{M}_m$ and $\mu = (m, 0)$:

$$\bar{L}_t(\mu_t) - \bar{L}_t(\mu) \leq \nabla \bar{L}_t(\mu_t)^T(\mu_t - \mu) = \sum_{j=1}^{d} \left[\frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)(m_{t,j} - m_j) + \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)\sigma_{t,j}\right].$$

Using the update formulas 10:

$$(m_{t+1,j} - m_j)^2 = (m_{t,j} - m_j)^2 + \eta_{t,j}^2 \sigma_{t,j}^4 \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)^2 - 2\eta_{t,j}\sigma_{t,j}^2 \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)(m_{t,j} - m_j)$$

and

$$\sigma_{t+1,j}^2 = \sigma_{t,j}^2 + \frac{\eta_{t,j}^2 \sigma_{t,j}^4}{2} \frac{\partial \bar{L}_t}{\partial \sigma\ _j}(m_{t,j}, \sigma_{t,j})^2 - \eta_{t,j}\sigma_{t,j}^2 \sqrt{1 + \left(\frac{\eta_{t,j}\sigma_{t,j}\frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)}{2}\right)^2} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)\sigma_{t,j}.$$

Rearranging the terms, we get:

$$\frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)(m_{t,j} - m_j) = \frac{(m_{t,j} - m_j)^2 - (m_{t+1,j} - m_j)^2}{2\eta_{t,j}\sigma_{t,j}^2} + \frac{\eta_{t,j}\sigma_{t,j}^2 \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)^2}{2}$$

and

$$\frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)\sigma_{t,j} = \frac{\sigma_{t,j}^2 - \sigma_{t+1,j}^2}{\eta_{t,j}\sigma_{t,j}^2 \sqrt{1 + \left(\frac{\eta_{t,j}\sigma_{t,j}\frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)}{2}\right)^2}} + \frac{\eta_{t,j}\sigma_{t,j}^2 \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)^2}{2\sqrt{1 + \left(\frac{\eta_{t,j}\sigma_{t,j}\frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)}{2}\right)^2}}.$$

We also use the boundedness of the gradients: for any $(m, \sigma) \in \mathcal{M}$, at any date $t$,

$$\sum_{j=1}^{d} \left[ \frac{\partial \bar{L}_t}{\partial m_j}(m,\sigma)^2 + \frac{\partial \bar{L}_t}{\partial \sigma_j}(m,\sigma)^2 \right] \leq L^2.$$

We upper bound the inverse of the square root by 1, the gradient by $L$ and we sum over time:

$$\begin{aligned}
\sum_{t=1}^{T} \bar{L}_t(\mu_t) - \bar{L}_t(\mu) &\leq \sum_{j=1}^{d}\sum_{t=1}^{T} \frac{(m_{t,j}-m_j)^2}{2}\left[\frac{1}{\eta_{t,j}\sigma_{t,j}^2} - \frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right] \\
&+ \sum_{j=1}^{d}\sum_{t=1}^{T} \frac{\eta_{t,j}\sigma_{t,j}^2}{2} \frac{\partial \bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2 \\
&+ \sum_{j=1}^{d}\sum_{t=1}^{T} \frac{\sigma_{t,j}^2}{2}\left[\frac{2}{\eta_{t,j}\sigma_{t,j}^2} - \frac{2}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right] \\
&+ \sum_{j=1}^{d}\sum_{t=1}^{T} \frac{\eta_{t,j}\sigma_{t,j}^2}{2} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t,\sigma_t)^2 \\
&= \sum_{j=1}^{d}\sum_{t=1}^{T} \frac{(m_{t,j}-m_j)^2}{2}\left[\frac{1}{\eta_{t,j}\sigma_{t,j}^2} - \frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right] \\
&+ \sum_{j=1}^{d}\sum_{t=1}^{T} \frac{\sigma_{t,j}^2}{2}\left[\frac{2}{\eta_{t,j}\sigma_{t,j}^2} - \frac{2}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right] \\
&+ \sum_{t=1}^{T} \frac{\eta_{t,j}\sigma_{t,j}^2}{2} \sum_{j=1}^{d}\left[\frac{\partial \bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2 + \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t,\sigma_t)^2\right] \\
&\leq \sum_{j=1}^{d}\sum_{t=1}^{T} \left[(m_{t,j}-m_j)^2 + \sigma_{t,j}^2\right]\left[\frac{1}{\eta_{t,j}\sigma_{t,j}^2} - \frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right] \\
&+ \sum_{t=1}^{T} \frac{\eta_{t,j}\sigma_{t,j}^2}{2} \sum_{j=1}^{d}\left[\frac{\partial \bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2 + \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t,\sigma_t)^2\right].
\end{aligned}$$

The key point in the following is that the difference

$$\frac{1}{\eta_{t,j}\sigma_{t,j}^2} - \frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2}$$

does not depend on $j$ on account of the formula $\eta_{t,j} = K/(\sqrt{t}\sigma_{t,j}^2) > 0$. We also recall that

$$\sum_{j=1}^{d} (m_{t,j}-m_j)^2 + \sigma_{t,j}^2 \leq D^2.$$

Moreover,

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T},$$

so setting $\eta_{t,j} = \frac{K}{\sqrt{t}\sigma_{t,j}^2} > 0$ with $K = \frac{D\sqrt{2}}{L}$, we finally have:

$$\begin{aligned}
\sum_{t=1}^{T} \bar{L}_t(\mu_t) - \bar{L}_t(\mu) &\leq \frac{1}{K} \sum_{t=1}^{T} (\sqrt{t} - \sqrt{t-1}) \sum_{j=1}^{d} [(m_{t,j} - m_j)^2 + \sigma_{t,j}^2] + \sum_{t=1}^{T} \frac{K}{\sqrt{t}} L^2 \\
&\leq \frac{D^2}{K} \sum_{t=1}^{T} (\sqrt{t} - \sqrt{t-1}) + \frac{KL^2}{2} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \\
&= \left( \frac{D^2}{K} + \frac{KL^2}{2} \right) \sqrt{T} \\
&= DL\sqrt{2T},
\end{aligned}$$

where $K$ is chosen so that it minimizes the bound.

**Strongly convex case:**

Now, we assume that each function $\bar{L}_t$ is $H$-strongly convex, for all $m \in \mathcal{M}_m$ and $\mu = (m, 0)$:

$$\begin{aligned}
\bar{L}_t(\mu_t) - \bar{L}_t(\mu) &\leq \nabla \bar{L}_t(\mu_t)^T (\mu_t - \mu) - \frac{H}{2} \|\mu_t - \mu\|^2 \\
&= \sum_{j=1}^{d} \left[ \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)(m_{t,j} - m_j) + \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)\sigma_{t,j} - \frac{H}{2}(m_{t,j} - m_j)^2 - \frac{H}{2}\sigma_{t,j}^2 \right].
\end{aligned}$$

Again,

$$\frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)(m_{t,j} - m_j) = \frac{(m_{t,j} - m_j)^2 - (m_{t+1,j} - m_j)^2}{2\eta_{t,j}\sigma_{t,j}^2} + \frac{\eta_{t,j}\sigma_{t,j}^2 \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)^2}{2}$$

and

$$\frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)\sigma_{t,j} = \frac{\sigma_{t,j}^2 - \sigma_{t+1,j}^2}{\eta_{t,j}\sigma_{t,j}^2 \sqrt{1 + \left( \frac{\eta_{t,j}\sigma_{t,j} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)}{2} \right)^2}} + \frac{\eta_{t,j}\sigma_{t,j}^2 \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)^2}{2\sqrt{1 + \left( \frac{\eta_{t,j}\sigma_{t,j} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)}{2} \right)^2}},$$

and then as previously with $\eta_{t,j} = \frac{2}{Ht\sigma_{t,j}^2}$:

$$\begin{aligned}
\sum_{t=1}^{T} \bar{L}_t(\mu_t) - \bar{L}_t(\mu) \leq &\sum_{j=1}^{d} \sum_{t=1}^{T} \frac{(m_{t,j} - m_j)^2}{2} \left[ \frac{1}{\eta_{t,j}\sigma_{t,j}^2} - \frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2} - H \right] \\
&+ \sum_{j=1}^{d} \sum_{t=1}^{T} \frac{\eta_{t,j}\sigma_{t,j}^2}{2} \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)^2 \\
&+ \sum_{j=1}^{d} \sum_{t=1}^{T} \frac{\sigma_{t,j}^2}{2} \left[ \frac{2}{\eta_{t,j}\sigma_{t,j}^2} - \frac{2}{\eta_{t-1,j}\sigma_{t-1,j}^2} - H \right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\eta_{t,j}\sigma_{t,j}^2}{2}\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)^2 \\
&\leq \sum_{j=1}^{d}\sum_{t=1}^{T}\frac{(m_{t,j}-m_j)^2}{2}\left[\frac{tH}{2}-\frac{(t-1)H}{2}-H\right] \\
&+\sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\sigma_{t,j}^2}{2}\left[tH-(t-1)H-H\right] \\
&+\sum_{t=1}^{T}\frac{1}{Ht}\sum_{j=1}^{d}\left[\frac{\partial\bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2+\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)^2\right] \\
&\leq \sum_{j=1}^{d}\sum_{t=1}^{T}\frac{(m_{t,j}-m_j)^2}{2}\left[\frac{H}{2}-H\right]+0+\sum_{t=1}^{T}\frac{L^2}{Ht} \\
&\leq \frac{L^2}{H}(1+\log(T)),
\end{aligned}
$$

which ends the proof. ■

**Proof** [Proof of Theorem 4] The proof is exactly the same as for Theorem 2. As previously, we first prove by recursion on $T$ that

$$
\forall\mu\in\mathcal{M},\ \sum_{t=1}^{T}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t)\leq\sum_{t=1}^{T}\mu^T\nabla\bar{L}_t(\mu_t)+\frac{\|\mu-\mu_1\|^2}{\eta}. \tag{24}
$$

It is obvious that it holds for $T=0$. Assume now that (24) holds for some integer $T-1$. Then for all $\mu\in M$,

$$
\begin{aligned}
\sum_{t=1}^{T}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t) &= \sum_{t=1}^{T-1}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t)+\mu_{T+1}^T\nabla\bar{L}_T(\mu_T) \\
&\leq \sum_{t=1}^{T-1}\mu^T\nabla\bar{L}_t(\mu_t)+\frac{\|\mu-\mu_1\|^2}{\eta}+\mu_{T+1}^T\nabla\bar{L}_T(\mu_T)
\end{aligned}
$$

as (24) holds for $T-1$. Apply this again to $\mu=\mu_{T+1}$:

$$
\begin{aligned}
\sum_{t=1}^{T}\mu_{t+1}^T\nabla\bar{L}_t(\mu_t) &\leq \sum_{t=1}^{T}\mu_{T+1}^T\nabla\bar{L}_t(\mu_t)+\frac{\|\mu-\mu_1\|^2}{\eta} \\
&= \min_{m\in\mathcal{M}}\left[\sum_{t=1}^{T}m^T\nabla\bar{L}_t(\mu_t)+\frac{\|\mu-\mu_1\|^2}{\eta}\right],\text{ by definition of }\mu_{T+1} \\
&\leq \sum_{t=1}^{T}\mu^T\nabla\bar{L}_t(\mu_t)+\frac{\|\mu-\mu_1\|^2}{\eta}
\end{aligned}
$$

for all $\mu\in\mathcal{M}$. Thus, (24) holds for $T$, and thus for integers.

We prove now that for any $t \in \mathbb{N}$,

$$\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \leq \eta L^2. \tag{25}$$

Indeed,

$$\begin{aligned} \mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) &= (\mu_t - \mu_{t+1})^T \nabla \bar{L}_t(\mu_t) \\ &\leq \|\mu_t - \mu_{t+1}\| \|\nabla \bar{L}_t(\mu_t)\| \\ &\leq L \|\mu_t - \mu_{t+1}\| \end{aligned} \tag{26}$$

as previously. Define

$$G_t(\mu) = \sum_{i=1}^{t-1} \mu^T \nabla \bar{L}_t(\mu_i) + \frac{\|\mu - \mu_1\|^2}{\eta}.$$

Obviously, $G_t$ is $1/\eta$-strongly convex: for any $(\mu, \mu')$,

$$G_t(\mu') - G_t(\mu) \geq (\mu' - \mu)^T \nabla G_t(\mu) + \frac{\|\mu' - \mu\|^2}{2\eta}.$$

In particular, $\mu_t$ is a minimizer of $G_t$:

$$G_t(\mu_{t+1}) - G_t(\mu_t) \geq \frac{\|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

Similarly,

$$G_{t+1}(\mu_t) - G_{t+1}(\mu_{t+1}) \geq \frac{\|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

Hence:

$$\bar{L}_t(\mu_t) - \bar{L}_t(\mu_{t+1}) \geq \frac{\|\mu_{t+1} - \mu_t\|^2}{\eta},$$

and then

$$\|\mu_{t+1} - \mu_t\| \leq \sqrt{\eta \left[\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t)\right]}$$

which combined with (26) leads to (25).

Finally, as for Theorem 2:

$$\sum_{t=1}^T \ell_t(\hat{\theta}_t) - \sum_{t=1}^T \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] \leq \eta T L^2 + \frac{\|\mu - \mu_1\|^2}{\eta},$$

which ends the proof. ■