Journal of Machine Learning Research 83 (2018) Submitted 10/27 2017; Published

Dimension-free Information Concentration via Exp-Concavity

Ya-Ping Hsieh YA-PING.HSIEH@EPFL.CH

Volkan Cevher VOLKAN.CEVHER@EPFL.CH

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
EPFL-STI-IEL-LIONS, Station 11
CH-1015 Lausanne, Switzerland

Editor:

Abstract

Information concentration of probability measures have important implications in learning theory. Recently, it is discovered that the information content of a log-concave distribution concentrates around their differential entropy, albeit with an unpleasant dependence on the ambient dimension. In this work, we prove that if the potentials of the log-concave distribution are *exp-concave*, which is a central notion for fast rates in online and statistical learning, then the concentration of information can be further improved to depend only on the exp-concavity parameter, and hence, it can be dimension independent. Central to our proof is a novel yet simple application of the variance Brascamp-Lieb inequality. In the context of learning theory, our concentration-of-information result immediately implies high-probability results to many of the previous bounds that only hold in expectation.

Keywords: Dimension-Free Concentration, Log-Concave Measures, Exp-Concavity, Variance Brascamp-Lieb Inequality, Differential Entropy

1. Introduction

We study the **information concentration** of probability measures: Given a probability density f and a random variable $X \sim f$, we ask how concentrated is the random variable $-\log f(X)$ around its mean $\mathbb{E}[-\log f(X)]$, which is simply the differential entropy of f.

We focus on the class of **log-concave** probability measures, whose densities are of the form $f(\mathbf{x}) \propto e^{-V(\mathbf{x})}$ for some convex function $V(\cdot)$. Information concentration for log-concave measures has found many applications in learning theory, ranging from aggregation (Dalalyan et al., 2016) and Bayesian decision theory (Pereyra, 2017), to, unsurprisingly, information theory (Raginsky et al., 2013). It also has immediate implications to online learning and PAC-Bayesian analysis (*cf.*, Section 4 for further discussion on this topic)

Bobkov et al. (2011) discovered the information concentration phenomenon for log-concave measures. Their result was later sharpened by Fradelizi et al. (2016), which establishes the current state-of-the-art. However, via the concentration bound in (Fradelizi et al., 2016), one can immediately notice a poor dependence on the dimension (see **Theorem 1**).

This unpleasant dependence is, however, not due to any deficit of the analysis: Even in the Gaussian case, the information concentration is known to be dimension-dependent

(Cover and Pombra, 1989), and the bound in (Fradelizi et al., 2016) matches the tightest known result. We can verify that the exponential distributions, another candidate for dimension-free concentration, share the same poor dimensional scaling.

Given these observations, one might pessimistically conjecture that no meaningful subclass of log-concave measures satisfies the information concentration in a dimension-free fashion. Hence, our main result comes as a surprise that, not only does there exist a large subclass of log-concave measures with dimension-free information concentration, but in addition, this subclass is extremely well-known to the machine learning community:

> **Our main result (informal statement):** *Let $f(\mathbf{x}) \propto e^{-V(\mathbf{x})}$ where V is η-**exp-concave**. Then, the information concentration of $f(\mathbf{x})$ solely depends on the exp-concavity parameter η, and not the ambient dimension.*

Many loss functions in machine learning are known to be exp-concave; a non-exhaustive list includes the squared loss, entropic loss, log-linear loss, SVMs with squared hinge loss, and log loss; see (Cesa-Bianchi et al., 2006) for more. Moreover, distributions of the type $e^{-V(\mathbf{x})}$, where V is exp-concave, appear frequently in many areas of learning theory. Consequently, our main result is tightly connected to learning with exp-concave losses; see Section 4.

Our main insight is that exp-concave functions are Lipschitz in a local norm, and log-concave measures satisfy the "Poincaré inequality in this local norm", namely the Brascamp-Lieb inequality. We elaborate more on the intuition in Appendix B.1. In retrospect, the proof of our main result is, once the right tools are identified, completely natural and elementary. In fact, our result basically implies that the exp-concavity arises naturally in the dimension-free information concentration.

The rest of the paper is organized as follows. We first set up notations and review basics of differential entropy and log-concave distributions in Section 2. In Section 3, which contains precise statements of the main result, we present various dimension-free inequalities for information concentration. We provide a counterexample to a conjecture, which is a natural strengthening of our results. We discuss implications of information concentration in Section 4 with motivating examples. The proofs of the main results can be found in Appendix B.

2. Preliminaries

2.1 Notations

For a function f, we write $\mathbb{E}_\mu f := \int f d\mu$ and $\mathrm{Var}_\mu(f) := \int f^2 d\mu - \left(\int f d\mu\right)^2$. We write $X \sim \mu$ for a random variable X associated with the probability measure μ.

In this paper, the norm $\|\cdot\|$ is always the Euclidean norm, and we use $\langle \cdot, \cdot \rangle$ for the Euclidean inner product. We use ∇V, $\nabla^2 V$, and ∂V to denote the gradient, Hessian, and subgraident of V, respectively. The notation $\mathcal{C}^k$ denotes the class of k-times differentiable functions with continuous k-th derivatives.

2.2 Differential Entropy and Log-Concave Distributions

Let μ be a probability measure having density f with respect to the Lebesgue measure and let $X \sim \mu$. The differential entropy (Cover and Thomas, 2012) of X is defined as

$$h(\mu) = h(X) \coloneqq \mathbb{E}_\mu[-\log f(X)]. \tag{1}$$

The random variable $\tilde{h}(\mu) = \tilde{h}(X) \coloneqq -\log f(X)$ is called the *information content* of μ.

We study the concentration of information content around the differential entropy:

$$\mathbb{P}\left(|\tilde{h}(X) - h(X)| > t\right) \le \alpha(t)$$

where $\alpha : \mathbb{R}^+ \to \mathbb{R}^+$ vanishes rapidly as t increases.

Throughout this paper, we consider *log-concave probability measures*, namely probability measures having density of the form

$$d\mu_V(\mathbf{x}) = \frac{e^{-V(\mathbf{x})}}{\int e^{-V}} d\mathbf{x}, \tag{2}$$

where V is a convex function such that $\int e^{-V} < \infty$. The function V is called the *potential* of the measure μ_V. For log-concave measures, the concentration of information content is equivalent to the concentration of the potential, i.e., $\mathbb{P}\left(|V - \mathbb{E}_{\mu_V} V| > t\right)$.

3. Dimension Free Concentration of Information for Exp-Concave Potentials

This section presents our main results.

We first review the state-of-the-art bound in Section 3.1. In Section 3.2, we demonstrate dimension-free information concentration when the underlying potential V is assumed to be exp-concave. All our results are of sub-exponential type; it is hence natural to ask if the sub-Gaussian counterparts are also true. We show that this is impossible even in dimension 1, by giving a counterexample in Section 3.3. Finally, we highlight some immediate consequences of our main results in Section 3.4. All proofs are deferred to Appendix B.

3.1 Previous Art

The state-of-the-art concentration bound for V is given by Fradelizi et al. (2016):

Theorem 1 (Information Concentration for Log-Concave Vectors) *Let $d\mu_V(\mathbf{x}) \coloneqq \frac{e^{-V(\mathbf{x})}}{\int e^{-V}} d\mathbf{x}$ be a d-dimensional log-concave probability measure. Then, we have*

1. $\mathrm{Var}(V(X)) \le d$.

2. *There exist universal constants c_1 and c_2 such that*

$$\mathbb{P}\left(|V - \mathbb{E}V| > t\right) \le c_1 \exp\left(-c_2 \min\left(t, \frac{t^2}{d}\right)\right). \tag{3}$$

This is the main result of (Fradelizi et al., 2016) combined with the well-known relation

$$t - \log(1+t) \simeq \min(t, t^2)$$

for every $t \geq 0$.

The bound (3) matches the tightest known results for $V = \frac{\|\cdot\|^2}{2}$ (i.e., the Gaussian case; see Cover and Pombra, 1989). However, notice that (3) has a poor dependence on the dimension d, as well as having the exponent being the worst case of t and $\frac{t^2}{d}$.

3.2 Our Results

We first recall the definition of exp-concave functions (Hazan, 2016):

Definition 2 *A function V is said to be η-exp-concave if $e^{-\eta V}$ is concave. Equivalently, V is η-exp-concave if the matrix inequality $\nabla^2 V(\mathbf{x}) \succeq \eta \nabla V(\mathbf{x}) \nabla V(\mathbf{x})^\top$ holds for all $\mathbf{x}$. Notice that an exp-concave function is necessarily convex.*

We next present three concentration inequalities for V in **Theorem 3-6**. **Theorem 3** serves as the prototype for all the concentration inequalities to come, however with restrictive conditions that severely limit its applicability. To overcome such dilemma, in **Theorem 5** and **6** we introduce practically motivated assumptions, and show how the restrictive conditions of **Theorem 3** can be removed without effecting the concentration.

3.2.1 Information Concentration: the Strictly Convex Case

The first main result of this paper is that, for $d\mu_V$ with V being η-exp-concave and strictly convex, the concentration of information content depends solely on η.

Theorem 3 *Assume that $V \in \mathcal{C}^2$ is η-exp-concave and $\nabla^2 V \succ 0$. Let $d\mu_V(\mathbf{x}) = \frac{e^{-V(\mathbf{x})}}{\int e^{-V}} d\mathbf{x}$ be the log-concave distribution associated with V. Then*

1. $\mathrm{Var}_{\mu_V}(V) \leq \frac{1}{\eta}$.
2. $\mathbb{P}\Big(|V - \mathbb{E}V| \geq t\Big) \leq 6 \exp\left(-\max(\sqrt{\eta}, \eta)t\right)$.

Notice that when $\frac{1}{\eta} \simeq d$, the bounds in **Theorem 1** and **3** yield comparable results. In this sense, $\frac{1}{\eta}$ can be viewed as the "effective dimension" regarding the concentration of information content.

3.2.2 Information Concentration: the General Convex Case

In many of the applications in learning theory (*cf.*, Section 4), the potential V is not guaranteed to be globally strictly convex. However, we have the following observation:

Lemma 4 *Assume that $V \in \mathcal{C}^2$ is η-exp-concave. Let $\mathcal{S}^+(\mathbf{x})$ be the subspace spanned by the eiganvectors corresponding to non-zero eigenvalues of $\nabla^2 V(\mathbf{x})$. Then $\nabla V(\mathbf{x}) \in \mathcal{S}^+(\mathbf{x})$ for all $\mathbf{x}$.*

Simply put, $\nabla^2 V$ may not be strictly convex in all directions, but it is always strictly convex in the direction of ∇V. Our second result shows that in this case, one can drop the global strict convexity of V while retaining exactly the same dimension-free concentration.

Theorem 5 *Assume that $V \in \mathcal{C}^2$ is η-exp-concave, but not necessarily strictly convex. Let $d\mu_V(\mathbf{x}) = \frac{e^{-V(\mathbf{x})}}{\int e^{-V}} d\mathbf{x}$ be the log-concave distribution associated with V. Then*

1. $\mathrm{Var}_{\mu_V}(V) \leq \frac{1}{\eta}$.

2. $\mathbb{P}\Big(|V - \mathbb{E}V| \geq t\Big) \leq 6 \exp\left(-\max(\sqrt{\eta}, \eta)t\right)$.

3.2.3 Information Concentration in the Presence of Nonsmooth Potential

The following case appears frequently in machine learning applications: The potential V can be decomposed as $V = V_1 + V_2$, where V_1 is a "nice" convex function (meaning satisfying either the assumptions in **Theorem 3** or **Theorem 5**), while V_2 is a nonsmooth convex function. Since V is neither differentiable nor strictly convex, results above do not apply.

Our third result is to show that, in this scenario, the term V_1 in fact enjoys dimension-free concentration as if the nonsmooth term V_2 is absent.

Theorem 6 *Let $V = V_1 + V_2$, where V_1 satisfies the assumptions in either **Theorem 3** or **Theorem 5**, and V_2 is a general convex function. Then we have*

$$\mathbb{P}\Big(|V_1 - \mathbb{E}V_1| \geq t\Big) \leq 6 \exp\left(-\max(\sqrt{\eta}, \eta)t\right), \tag{4}$$

where the probability is with respect to the total measure $d\mu_V$, and η is the exp-concavity parameter of V_1.

3.3 A Counterexample to Sub-Gaussian Concentration of Information Content

So far, we have established dimension-free concentration of sub-exponential type under various conditions. An ansatz is whether under the same assumptions, one has dimension-free *sub-Gaussian concentration*; i.e., a deviation inequality of the form

$$\mathbb{P}\left(|V - \mathbb{E}V| \geq t\right) \leq c_1 e^{-c(\eta)t^2} \tag{5}$$

for a universal constant c_1 and some constant $c(\eta)$ depending only on η.

In this subsection, we provide a counterexample to this conjecture, showing that this is impossible even in dimension 1.

Consider the one-dimensional case where $V(x) = -\log x$ and the support is $\Omega := (0,1)$. Notice that V is trivially 1-exp-concave. If (5) holds for V, then we would have

$$\begin{aligned}
\mathbb{E}e^{\lambda(V-\mathbb{E}V)^2} &= \int_0^\infty \mathbb{P}\left(e^{\lambda(V-\mathbb{E}V)^2} > x\right) dx \\
&\leq 2\lambda \int_0^\infty c_1 e^{-c(\eta)t^2} t e^{\lambda t^2} dt \\
&< \infty
\end{aligned}$$

if $\lambda < c(\eta)$. However, a straightforward computation shows that

$$\mathbb{E}e^{\lambda(V-\mathbb{E}V)^2} = \frac{e^{\frac{\lambda}{64}}}{2}\int_0^1 x^{1.25}e^{\lambda(\log x)^2}dx = \infty \tag{6}$$

for every $\lambda > 0$. We hence cannot have any sub-Gaussian concentration for V.

It is easy to generalize this example to any dimension.

3.4 Immediate Consequences

An immediate consequence of information concentration is that many important densities in information theory also concentrate.

Corollary 7 (Concentration of Information Densities) *Let $f(\mathbf{x},\mathbf{y}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a joint log-concave density of the random variable pair (X,Y). Denote the marginal distribution of the first argument by $f(\mathbf{x}) := \int_{\mathbb{R}^d} f(\mathbf{x},\mathbf{y})d\mathbf{y}$ and similarly for $f(\mathbf{y})$, and denote the conditional distribution by $f(\mathbf{y}|\mathbf{x}) := \frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{x})}$. Then there exist universal constants c_1, c_2 such that the following holds:*

1. $\mathbb{P}\left(|-\log f(Y|X) - \mathbb{E}[-\log f(Y|X)]| > t\right) \leq 2c_1 \exp\left(-\frac{c_2}{2}\min\left(t, \frac{t^2}{d}\right)\right)$.
2. $\mathbb{P}\left(\left|-\log\frac{f(X,Y)}{f(X)f(Y)} - \mathbb{E}\left[-\log\frac{f(X,Y)}{f(X)f(Y)}\right]\right| > t\right) \leq 3c_1 \exp\left(-\frac{c_2}{3}\min\left(t, \frac{t^2}{d}\right)\right)$.

If, in addition, that $-\log f(\mathbf{x},\mathbf{y})$, $-\log f(\mathbf{x})$, and $-\log f(\mathbf{y})$ are η-exp-concave and $-\log f(\cdot,\cdot)$ is strictly convex. Then the exponents in the above bounds can be improved to $\max(\sqrt{\eta},\eta)t$.

Notice that $h(Y|X) := \mathbb{E}\left[-\log f(Y|X)\right]$ is the *conditional (differential) entropy*, and $I(X;Y) := \mathbb{E}\left[-\log\frac{f(X,Y)}{f(X)f(Y)}\right]$ is the mutual information. The (random) quantities $-\log f(Y|X)$ and $-\log\frac{f(X,Y)}{f(X)f(Y)}$ play prominent roles in recent advances of non-asymptotic information theory; see Polyanskiy (2010) and the references therein.

Proof A celebrated result of Prékopa (1971) states that the marginals of log-concave measures are also log-concave. The corollary then follows by the well-known decomposition $h(Y|X) = h(X,Y) - h(X)$ and $I(X;Y) = h(X) + h(Y) - h(X,Y)$. ■

4. Motivating Examples

Unsurprisingly, information concentration has many applications in learning theory; we present three examples in this section. To avoid lengthy but straightforward calculations, we shall omit the details and refer the readers to proper literature.

Below, we consider loss functions of the form $L_n(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \ell_i(\mathbf{x})$, where ℓ_i's are exp-concave. By **Lemma 10** in Appendix A, the total loss L_n is also exp-concave. Denote the exp-concave parameter of L_n by η.

We remark that, in general, η can depend on the dimension d or the sample size n. A comparison of the favorable regimes for different η's is presented in Table 1.

	Fradelizi et al. (2016)	Ours, $\eta = \Omega(1)$	Ours, $\eta = \Omega\left(\frac{1}{d}\right)$
$t = \Theta(1)$	$\exp\left(-\frac{1}{d}\right)$	$\exp(-1)$	$\exp\left(-\frac{1}{\sqrt{d}}\right)$
$t = \Theta(\sqrt{d})$	$\exp(-1)$	$\exp\left(-\sqrt{d}\right)$	$\exp(-1)$
$t = \Theta(d)$	$\exp(-d)$	$\exp(-d)$	$\exp\left(-\sqrt{d}\right)$

Table 1: The deviation $\mathbb{P}\left(\left|V - \mathbb{E}V\right| > t\right)$ dictated by different concentration inequalities.

4.1 High-Probability Regret Bounds for Exponential Weight Algorithms

Exp-concave losses have received substantial attention in online learning as they exhibit logarithmic regret (Hazan et al., 2007). One class of algorithms attaining logarithmic regret is based on the *Exponential Weight*, which makes prediction according to

$$\mathbf{x}_{t+1} = \mathbb{E}_{\pi_t} X, \tag{7}$$

where

$$\pi_t(\mathbf{x}) \propto e^{-nL_n(\mathbf{x})}. \tag{8}$$

A common belief is that the algorithm (7) is inefficient to implement, and practitioners would more opt into first-order methods such as the (see Hazan et al., 2007) Online Newton Step (which is also somewhat inefficient: every iteration requires inverting a matrix and a projection). However, recent years have witnessed a surge of interest in the sampling schemes, mainly due to its connection to the ultra-simple Stochastic Gradient Descent (Welling and Teh, 2011). Theoretical (Bubeck et al., 2015; Durmus and Moulines, 2016; Dalalyan, 2017; Dalalyan and Karagulyan, 2017; Cheng and Bartlett, 2018) and empirical (Welling and Teh, 2011; Ahn et al., 2012; Rezende et al., 2014; Blei et al., 2017) studies of sampling schemes have now become one of the most active areas in machine learning.

In view of these recent developments, it is natural to consider, instead of the expected prediction (7), taking samples $X_{t,1}, X_{t,2}, ..., X_{t,N} \sim \pi_t$ and predict $\bar{X}_t := \frac{1}{N}\sum_{i=1}^N X_{t,i}$. The following corollary of our main result establishes the desirable concentration property of $\bar{X}_t$.

Corollary 8 *Let $\{X_i\}_{i=1}^N$ be i.i.d. samples from the distribution $\frac{e^{-V}}{\int e^{-V}}$. Assume that V satisfies either the assumptions of **Theorem 3** or **Theorem 5**. Then*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N V(X_i) - \mathbb{E}V\right| > t\right) \leq 2e^{-N(\sqrt{\eta}t - \log 3)}. \tag{9}$$

Proof For simplicity, assume $\eta = 1$; the general case is similar.

By the classic Chernoff bounding technique, we can compute

$$\begin{aligned}\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} V(X_i) - \mathbb{E}V > t\right) &= \mathbb{P}\left(e^{\sum_{i=1}^{N}(V(X_i)-\mathbb{E}V)} > e^{Nt}\right) \\ &\leq e^{-Nt}\left(\mathbb{E}e^{(V(X)-\mathbb{E}V)}\right)^N \\ &\leq e^{-Nt}\cdot 3^N \\ &= e^{-N(t-\log 3)},\end{aligned}$$

where the second inequality follows from (28) with $\eta = 1$. ■

Plugging (9) into the expected regret bounds for the Exponential Weight algorithm (e.g., Hazan et al., 2007), we immediately obtain high-probability regret bounds.

Similar arguments hold for random walk-based approaches in online learning (Narayanan and Rakhlin, 2010).

4.2 Posterior Concentration of Bayesian and Pac-Bayesian Analysis

The (pseudo-)posterior distribution plays a fundamental role in the PAC-Bayesian theory:

$$\hat{\pi}(\mathbf{x}) \propto \exp\left(-nV_n(\mathbf{x})\right), \tag{10}$$

where $V_n(\mathbf{x}) = L_n(\mathbf{x}) - \frac{1}{n}\log \pi_0(\mathbf{x})$. Here, $\mathbf{x}$ represents the parameter vector and π_0 is the prior distribution. It is well-known that (10) is optimal in PAC-Bayesian bounds for the expected (over the posterior distribution on the parameter set) population risk (Catoni, 2007). Moreover, when the loss functions ℓ_i's are the negative log-likelihood of the data, the optimal PAC-Bayesian posterior (10) coincides with the Bayesian posterior; see (Zhang, 2006) or the more recent (Germain et al., 2016).

We now consider the high-probability bound in the following sense: Instead of taking the expectation over $\hat{\pi}$ as previously done, we draw a random sample $X \sim \hat{\pi}$, and ask what is the population risk for X. Besides its apparent theoretical interest, such characterization is also important in practice, as there exist many sampling schemes for log-concave distributions $\hat{\pi}$ (Lovász and Vempala, 2007; Bubeck et al., 2015; Durmus and Moulines, 2016; Dalalyan, 2017), while computing the mean is in general costly (the mean is typically obtained through a large amount of sampling anyway).

A straightforward application of **Theorem 6** shows that, if the prior π_0 is log-concave, then $L_n(X)$ concentrates around $\mathbb{E}_{\hat{\pi}}L_n(X)$; notice that many popular priors (uniform, Gaussian, Laplace, etc.) are log-concave. On the other hand, concentration of the empirical risk L_n around the population risk is a classical theme in statistical learning. To conclude, **Theorem 6** implies high-probability results for the PAC-Bayesian bounds. In view of the equivalence established in (Germain et al., 2016), we also obtain concentration for the Bayesian posterior in the case of negative log-likelihood loss.

4.3 Bayesian Highest Posterior Density Region

Let $\hat{\pi}$ be the posterior distribution as in (10). In Bayesian decision theory, the optimal confidence region associated with a level α is given by the Highest Posterior Density (HPD) region (Robert, 2007), which is defined as

$$C_\alpha^\star \coloneqq \{\mathbf{x} \in \mathbb{R}^d \mid V_n(\mathbf{x}) \leq \gamma_\alpha\} \tag{11}$$

where γ_α is chosen so that $\int_{C_\alpha^\star} \hat{\pi}(\mathbf{x}) d\mathbf{x} = 1 - \alpha$.

Using concentration of the information content for log-concave distributions, Pereyra (2017) showed that $C_\alpha^\star$ is contained in the set

$$\tilde{C}_\alpha \coloneqq \{\mathbf{x} \in \mathbb{R}^d \mid V_n(\mathbf{x}) \leq V_n(\mathbf{x}^\star) + dt_\alpha + d\}, \tag{12}$$

where $\mathbf{x}^\star \coloneqq \arg\max_{\mathbf{x}} V_n(\mathbf{x})$ is the MAP parameter, and $t_\alpha = c_1\sqrt{\frac{\log(1/\alpha)}{d}}$ for some constant c_1. A straightforward application of our results shows that, when the data term L_n in V_n is η-exp-concave, then we can improve (12). For simplicity, let us focus on the uniform prior ($\pi_0 =$ constant). Adapting the analysis in Pereyra (2017), we can show that $C_\alpha^\star$ is contained in the set

$$\tilde{C}_\alpha^\eta \coloneqq \{\mathbf{x} \in \mathbb{R}^d \mid V_n(\mathbf{x}) \leq V_n(\mathbf{x}^\star) + t_\alpha^\eta + d\}, \tag{13}$$

where $t_\alpha^\eta = c_2 \log(1/\alpha) \cdot \sqrt{\frac{n}{\eta}}$. Comparing (12) and (13), we see that (ignoring logarithmic terms) we get improvements whenever $\eta = \Omega\left(\frac{n}{d}\right)$. This is typically the case in high-dimensional statistics (Bühlmann and Van De Geer, 2011) or compressive sensing (Ji et al., 2008; Foucart and Rauhut, 2013) where $n \ll d$.

Similar results can be established for the Gaussian and Laplace prior, where one can invoke results in (Cover and Pombra, 1989) and (Talagrand, 1995) to deduce the concentration of the prior term. We omit the details.

5. Conclusion

We have shown that for log-concave distributions with exp-concave potentials, the information concentration is dictated by its exp-concavity parameter η. Information theoretically speaking, η (or rather $\frac{1}{\eta}$) can be viewed as some sort of effective dimension, in the sense that $\frac{1}{\eta}$ and d play very similar roles in both the variance and concentration controls, the former for log-concave measures with exp-concave potential and the latter for general log-concave measures. Such a understanding enables us to derive high-probability results for many of the machine learning algorithms, including the Bayesian, PAC-Bayesian, and Exponential Weight type approaches.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^o 725594 - time-data).

References

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1771–1778, 2012.

David Alonso-Gutiérrez and Jesús Bastero. *Approaching the Kannan-Lovász-Simonovits and variance conjectures*, volume 2131. Springer, 2015.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.

Sergey Bobkov, Mokshay Madiman, et al. Concentration of the information in data with log-concave distributions. *The Annals of Probability*, 39(4):1528–1543, 2011.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *arXiv preprint arXiv:1507.02564*, 2015.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Olivier Catoni. *PAC-Bayesian Supervised Classification: the Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007.

Nicolo Cesa-Bianchi, Gabor Lugosi, and Learning Prediction. Games, 2006.

Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. *Proceedings of Machine Learning Research*, 83, 2018.

Dario Cordero-Erausquin. Transport inequalities for log-concave measures, quantitative forms and applications. *Canadian Journal of Mathematics*, 69:481–501, 2017.

Thomas M Cover and Sandeep Pombra. Gaussian feedback capacity. *IEEE Transactions on Information Theory*, 35(1):37–43, 1989.

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

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

Arnak S Dalalyan, Edwin Grappin, and Quentin Paris. On the exponentially weighted aggregate with the laplace prior. *arXiv preprint arXiv:1611.08483*, 2016.

Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. 2016.

Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

Matthieu Fradelizi, Mokshay Madiman, and Liyao Wang. Optimal concentration of information content for log-concave densities. In *High Dimensional Probability VII*, pages 45–60. Springer, 2016.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

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

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

Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(1):541–559, 1995.

Michel Ledoux. Spectral gap, logarithmic sobolev constant, and geometric bounds. *Surveys in differential geometry*, 9:219–240, 2004.

Michel Ledoux. *The concentration of measure phenomenon.* American Mathematical Soc., 2005.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Hariharan Narayanan and Alexander Rakhlin. Random walk approach to regret minimization. In *Advances in Neural Information Processing Systems*, pages 1777–1785, 2010.

Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Marcelo Pereyra. Maximum-a-posteriori estimation with bayesian confidence regions. *SIAM Journal on Imaging Sciences*, 10(1):285–302, 2017.

Yury Polyanskiy. *Channel coding: non-asymptotic fundamental limits.* Princeton University, 2010.

András Prékopa. Logarithmic concave measures with applications. *Acta Sci. Math*, 32: 301–316, 1971.

Maxim Raginsky, Igal Sason, et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Science & Business Media, 2007.

Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathematiques de l'IHES*, 81(1):73–205, 1995.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Appendix A. Properties of Exp-Concave Functions

We present two useful properties of exp-concave functions in this appendix. While these properties are well-known to the experts, we provide the proofs for completeness.

Lemma 9 *Assume that V is η-exp-concave and $\nabla^2 V \succ 0$. Then we have*

$$\left\langle \nabla^2 V^{-1}(\mathbf{x}) \nabla V(\mathbf{x}), \nabla V(\mathbf{x}) \right\rangle \leq \frac{1}{\eta} \tag{14}$$

for all $\mathbf{x}$.

Proof Since V is η-exp-concave, we have

$$\frac{1}{\eta}\nabla^2 V \succeq \nabla V \nabla V^\top. \tag{15}$$

Let $v = \frac{\nabla V}{\|\nabla V\|}$ and $R \coloneqq I - vv^\top$. For any $\delta > 0$, we have

$$\begin{aligned} \frac{1}{\eta}\nabla^2 V + \delta I &\succeq \frac{1}{\eta}\nabla^2 V + \delta R \\ &\succeq \nabla V \nabla V^\top + \delta R \\ &= \left(\|\nabla V\|^2 - \delta\right) vv^\top + \delta I \\ &\succ 0 \end{aligned} \tag{16}$$

if $\delta < \|\nabla V\|^2$. Using the fact that $B \succeq A \succ 0$ implies $A^{-1} \succeq B^{-1} \succ 0$, we get

$$\left(\frac{1}{\eta}\nabla^2 V + \delta I\right)^{-1} \preceq \frac{1}{\delta}\left(I + tvv^\top\right)^{-1} \tag{17}$$

where $t \coloneqq \frac{\|\nabla V\|^2}{\delta} - 1$. The Sherman–Morrison formula implies

$$\left(I + tvv^\top\right)^{-1} = I - \frac{t}{1+t} vv^\top, \tag{18}$$

and hence,

$$\begin{aligned}\left\langle \left(\frac{1}{\eta}\nabla^2 V + \delta I\right)^{-1} \nabla V, \nabla V \right\rangle &\leq \frac{1}{\delta}\left(\|\nabla V\|^2 - \frac{t}{1+t}\|\nabla V\|^2\right) \\ &= \frac{1+t}{\|\nabla V\|^2}\left(\|\nabla V\|^2 - \frac{t}{1+t}\|\nabla V\|^2\right) \\ &= 1. \end{aligned} \tag{19}$$

On the other hand, since $\nabla^2 V \succ 0$, we have

$$\begin{aligned}\lim_{\delta\to 0}\left\langle \left(\frac{1}{\eta}\nabla^2 V + \delta I\right)^{-1} \nabla V, \nabla V \right\rangle &= \left\langle \left(\frac{1}{\eta}\nabla^2 V\right)^{-1} \nabla V, \nabla V \right\rangle \\ &= \eta \left\langle \nabla^2 V^{-1} \nabla V, \nabla V \right\rangle. \end{aligned}$$

The proof is hence completed by letting $\delta \to 0$ in (19). ■

Lemma 10 *Let V_i's be η_i-exp-concave functions for $i = 1, 2, ..., k$. Then $\sum_{i=1}^k V_i$ is η-exp-concave with $\frac{1}{\eta} = \sum_{i=1}^k \frac{1}{\eta_i}$.*

Proof Let X be any random variable. Using the exp-concavity and Hölder's inequality, we get

$$\begin{aligned} e^{-\eta(V_1+V_2)(\mathbb{E}X)} &\geq \left(\mathbb{E}e^{-\eta_1 V_1(X)}\right)^{\frac{\eta}{\eta_1}} \cdot \left(\mathbb{E}e^{-\eta_2 V_2(X)}\right)^{\frac{\eta}{\eta_2}} \\ &= \|e^{-V_1}\|_{\eta_1}^{\eta} \cdot \|e^{-V_2}\|_{\eta_2}^{\eta} \\ &\geq \|e^{-(V_1+V_2)}\|_{\eta}^{\eta} \\ &= \mathbb{E}e^{-\eta\left(V_1(X)+V_2(X)\right)}. \end{aligned}$$

Here, $\|e^{-V_1}\|_{\eta_1} := \left(\mathbb{E}e^{-\eta_1 V_1}\right)^{\frac{1}{\eta_1}}$ and similarly for $\|e^{-V_2}\|_{\eta_2}$.

The general case follows from induction. ■

Appendix B. Proofs of the Main Results

We prove the main results in this appendix. Our analysis crucially relies on the *variance Brascamp-Lieb inequality*, recalled and elaborated in Appendix B.1. Appendix B.2-4 are devoted to the proofs of **Theorem 3-6**, respectively.

B.1 Proof Ideas

For a probability measure μ, we say that μ satisfies the Poincaré inequality with constant λ_1 if

$$\lambda_1 \mathrm{Var}_\mu(f) \leq \int \|\nabla f\|^2 d\mu \tag{20}$$

for all locally Lipschitz f. It is well-known that if (20) is satisfied for μ, then all the Lipschitz functions concentrate exponentially (Ledoux, 2004, 2005):

$$\forall\ 1-\text{Lipschitz } f, \quad \mathbb{P}\left(|f-\mathbb{E}f|>t\right)\leq c_1 e^{-\sqrt{\lambda_1}t} \tag{21}$$

for some universal constant c_1.

At first glance, our theorems seem to have little to do with the Poincaré inequality, since

1. It is not known whether a log-concave distribution satisfies the Poincaré inequality with a dimension-independent constant (this is the content of the Kannan-Lovász-Simonovits conjecture; see Kannan et al., 1995; Alonso-Gutiérrez and Bastero, 2015).

2. Typically, the potential V is not Lipschitz (consider the Gaussian distribution where $V(\mathbf{x})=\frac{\|\mathbf{x}\|^2}{2}$). Moreover, even if V is Lipschitz, the Lipschitz constant often depends on the dimension (consider the exponential distribution where $\|\nabla V\|=\Theta(\sqrt{d})$).

The important observation in this paper is that the appropriate norm in (20) for information concentration is not the Euclidean norm (or any ℓ_p-norm), but instead the (dual of the) *local norm* defined by the potential V itself, namely $\|\mathbf{y}\|_{\mathbf{x}}\coloneqq\left\langle\nabla^2V(\mathbf{x})\mathbf{y},\mathbf{y}\right\rangle$.

Lemma 9 in Appendix A expresses the fact that η-exp-concave functions are Lipschitz with respect to this local norm, and the *Brascamp-Lieb inequality* below provides a suitable strengthening of the Poincaré inequality:

Theorem 11 (Brascamp-Lieb Inequality) *Let $d\mu_V(\mathbf{x})=\frac{e^{-V(\mathbf{x})}}{\int e^{-V}}d\mathbf{x}$ be a log-concave probability measure with $V\in\mathcal{C}^2$ and $\nabla^2V\succ 0$. Then for all locally Lipschitz function $f\in L_2(\mu_V)$, we have*

$$\mathrm{Var}_{\mu_V}(f)\leq\int\left\langle\nabla^2V^{-1}\nabla f,\nabla f\right\rangle d\mu_V. \tag{22}$$

We shall see that the Brascamp-Lieb inequality provides precisely the desired control of the Lipschitzness of V in terms of the aforementioned dual local norm. Once this is observed, the rest of the proof is a routine in deducing from Poincaré inequality the sub-exponential concentration of Lipschitz functions.

We remark that our approach is, in retrospect, completely natural and elementary. However, to the best of our knowledge, our work is the first to combine the Brascamp-Lieb inequality (22) with the local norm of the form $\|\mathbf{y}\|_{\mathbf{x}}\coloneqq\left\langle\nabla^2V(\mathbf{x})\mathbf{y},\mathbf{y}\right\rangle$.

B.2 Proof of Theorem 3

The first assertion is a simple application of the Brascamp-Lieb inequality (22) and **Lemma 9**.

We now prove the concentration inequality. We first show that $\mathbb{P}\Big(|V-\mathbb{E}V|\geq t\Big)\leq 6\exp\left(-\sqrt{\eta}t\right)$. Applying (22) to $f=\exp\left(\frac{\lambda(V-\mathbb{E}V)}{2}\right)$, we get

$$\begin{aligned}\mathrm{Var}_{\mu_V}(f)&\leq\frac{\lambda^2}{4}\int f^2\left\langle\nabla^2V^{-1}\nabla V,\nabla V\right\rangle d\mu_V\\&\leq\frac{\lambda^2}{4\eta}\int f^2 d\mu_V\end{aligned} \tag{23}$$

by **Lemma 9**. Let $M(\lambda) \coloneqq \mathbb{E}\exp\left(\lambda(V - \mathbb{E}V)\right)$. Then the inequality (23) reads

$$M(\lambda) - M\left(\frac{\lambda}{2}\right)^2 \leq \frac{\lambda^2}{4\eta} M(\lambda), \tag{24}$$

and hence

$$M(\lambda) \leq \frac{1}{1 - \frac{\lambda^2}{4\eta}} M\left(\frac{\lambda}{2}\right)^2. \tag{25}$$

Apply (25) recursively to obtain

$$M(\lambda) \leq \Pi_{k=1}^{K-1} \left(\frac{1}{1 - \frac{\lambda^2}{4^{k+1}\eta}}\right)^{2^k} M\left(\frac{\lambda}{2^K}\right)^{2^K}. \tag{26}$$

Since $M(\lambda) = 1 + o(\lambda)$, we have

$$M\left(\frac{\lambda}{2^K}\right)^{2^K} = \left(1 + o\left(\frac{\lambda}{2^K}\right)\right)^{2^K} \to 1$$

as $K \to \infty$. Hence (26) implies

$$M(\lambda) \leq \Pi_{k=1}^{\infty} \left(\frac{1}{1 - \frac{\lambda^2}{4^{k+1}\eta}}\right)^{2^k}, \tag{27}$$

which in turn gives

$$M(\sqrt{\eta}) \leq 3. \tag{28}$$

The proof can now be completed by the classic Chernoff bounding technique:

$$\begin{aligned} \mathbb{P}\left(V - \mathbb{E}V \geq t\right) &= \mathbb{P}\left(e^{\sqrt{\eta}(V - \mathbb{E}V)} \geq e^{\sqrt{\eta}t}\right) \\ &\leq e^{-\sqrt{\eta}t} M(\sqrt{\eta}) \\ &\leq 3e^{-\sqrt{\eta}t}. \end{aligned} \tag{29}$$

Now, the inequality (29) implies that for any 1-exp-concave V, we have

$$\mathbb{P}\left(V - \mathbb{E}V \geq t\right) \leq 3e^{-t}.$$

If V is η-exp-concave, ηV is 1-exp-concave, and hence we conclude that

$$\mathbb{P}\left(V - \mathbb{E}V \geq \frac{t}{\eta}\right) \leq 3e^{-t};$$

that is to say,

$$\mathbb{P}\left(V - \mathbb{E}V \geq t\right) \leq 3e^{-\eta t}. \tag{30}$$

The bound for $\mathbb{P}\left(V - \mathbb{E}V \leq -t\right)$ is similar.

The proof is completed by taking the best case of (29) and (30), and applying the union bound.

B.3 Proof of Lemma 4 and Theorem 5

We first prove **Lemma 4**.

For any point $\mathbf{x}$, let $\{\mathbf{a}_i\}_{i=1}^k$ be an orthonormal basis for $\mathcal{S}^+(\mathbf{x})$, assumed to have dimension k. We extend $\{\mathbf{a}_i\}_{i=1}^k$ to an orthonormal basis for $\mathbb{R}^d$ as $\{\mathbf{a}_i\}_{i=1}^d$, and we decompose $\nabla V(\mathbf{x}) = \sum_{i=1}^d c_i \mathbf{a}_i$ for some real numbers c_i's.

For the purpose of contradiction, assume that $\nabla V(\mathbf{x}) \notin \mathcal{S}^+(\mathbf{x})$. Then $c_j \neq 0$ for some $j \in \{k+1, k+2, ..., d\}$. But then

$$\left\langle \nabla^2 V(\mathbf{x}) \mathbf{a}_j, \mathbf{a}_j \right\rangle = 0$$

while

$$\mathbf{a}_j^\top \nabla V(\mathbf{x}) \nabla V(\mathbf{x})^\top \mathbf{a}_j = c_j^2 > 0,$$

contradicting the exp-concavity of V. This finishes the proof of **Lemma 4**.

We now turn to **Theorem 5**.

Let $\epsilon > 0$ be arbitrarily small, and consider the quantity

$$\left\langle \left(\nabla^2 V + \epsilon I\right)^{-1} \nabla V, \nabla V \right\rangle. \tag{31}$$

Lemma 4 implies that (31) is equal to

$$\left\langle \left(\nabla^2 V + \epsilon I_{\mathcal{S}^+}\right)^{-1} \nabla V, \nabla V \right\rangle, \tag{32}$$

where $I_{\mathcal{S}^+}$ is the identity map on the subspace $\mathcal{S}^+$. Since V is strictly convex restricted to $\mathcal{S}^+$, and since $\nabla V \in \mathcal{S}^+$, **Lemma 9** then implies

$$\left\langle \left(\nabla^2 V + \epsilon I\right)^{-1} \nabla V, \nabla V \right\rangle \leq \frac{1}{\eta} \tag{33}$$

for all $\epsilon > 0$.

Consider $\tilde{V} = V + \frac{\epsilon \|\cdot\|^2}{2}$, and let $f = \exp\left(\frac{\lambda(V - \mathbb{E}V)}{2}\right)$. Since $\tilde{V}$ is strictly convex, we may invoke the Brascamp-Lieb inequality (22) to conclude

$$\begin{aligned} \mathrm{Var}_{\mu_{\tilde{V}}}(f) &\leq \frac{\lambda^2}{4} \int f^2 \left\langle \nabla^2 \tilde{V}^{-1} \nabla V, \nabla V \right\rangle d\mu_{\tilde{V}} \\ &\leq \frac{\lambda^2}{4\eta} \int f^2 d\mu_{\tilde{V}} \end{aligned} \tag{34}$$

where the second inequality follows from (33). Letting $\epsilon \to 0$ in (34) then gives

$$\mathrm{Var}_{\mu_V}(f) \leq \frac{\lambda^2}{4\eta} \int f^2 d\mu_V. \tag{35}$$

The rest of the proof is similar to that of **Theorem 3**; we omit the details.

B.4 Proof of Theorem 6

We will need the following strengthened Brascamp-Lieb inequality, which might be of independent interest. Once the following theorem is established, one can follow a similar proof as in Appendix B.2. We omit the details, and focus on the proof of the following theorem in the rest of this appendix.

Theorem 12 (Nonsmooth Brascamp-Lieb Inequality) *Let $d\mu_{\tilde{V}}(\mathbf{x}) = \frac{e^{-\tilde{V}(\mathbf{x})}}{\int e^{-\tilde{V}}} d\mathbf{x}$ be a log-concave measure with $\tilde{V} = V + U$, where $V \in \mathcal{C}^2$, $\nabla^2 V \succ 0$, and U is convex but possibly non-differentiable. Then for all locally Lipschitz function $f \in L_2(\mu_{\tilde{V}})$, we have*

$$\mathrm{Var}_{\mu_{\tilde{V}}}(f) \leq \int \left\langle \nabla^2 V^{-1} \nabla f, \nabla f \right\rangle d\mu_{\tilde{V}}. \tag{36}$$

Proof [Proof of **Theorem 12**] Define the cost functions

$$c_{\tilde{V}}(\mathbf{x}, \mathbf{y}) := \tilde{V}(\mathbf{y}) - \tilde{V}(\mathbf{x}) - \left\langle \partial \tilde{V}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle \tag{37}$$

and

$$c_V(\mathbf{x}, \mathbf{y}) := V(\mathbf{y}) - V(\mathbf{x}) - \left\langle \nabla V(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle. \tag{38}$$

By **Proposition 1.1** of Cordero-Erausquin (2017) (see also p.482 for the non-differentiable case), we know that the measure $d\mu_{\tilde{V}}$ satisfies the transportation cost inequality:

$$\mathcal{W}_{c_{\tilde{V}}}(\mu_{\tilde{V}}, \nu) \leq D(\nu \| \mu_{\tilde{V}})$$

for any probability measure ν. Here, $\mathcal{W}_{c_{\tilde{V}}}(\mu_{\tilde{V}}, \nu) := \inf_{X,Y} \mathbb{E} c_{\tilde{V}}(X, Y)$ where the infimum is over all joint distributions with marginals $X \sim \mu_{\tilde{V}}$ and $Y \sim \nu$, and $D(\nu \| \mu_{\tilde{V}})$ is the relative entropy between ν and $\mu_{\tilde{V}}$. Since U is convex, we have $c_{\tilde{V}}(\mathbf{x}, \mathbf{y}) \geq c_V(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y}$, and hence $\mu_{\tilde{V}}$ satisfies the weaker transportation cost inequality

$$\mathcal{W}_{c_V}(\mu_{\tilde{V}}, \nu) \leq D(\nu \| \mu_{\tilde{V}}). \tag{39}$$

The theorem can then be deduced from a standard linearization procedure that is well-known since the classic (Otto and Villani, 2000). The rest of the proof below is a suitable adaptation of the version in (Cordero-Erausquin, 2017).

Since continuous functions with compact support are dense in $L_2(\mu_{\tilde{V}})$, we will prove **Theorem 12** for any continuous function with compact support. Notice that such functions are necessarily Lipschitz and hence differentiable $\mu_{\tilde{V}}$-almost everywhere. Since modifying f in a set of $\mu_{\tilde{V}}$-measure 0 does not effect (36), we may henceforth assume that $f \in \mathcal{C}^1$ and has compact support.

Since $V \in \mathcal{C}^2$, $\nabla^2 V(\mathbf{y})$ is uniformly continuous on any compact set, and hence we have

$$c_V(\mathbf{y} + \mathbf{h}, \mathbf{y}) = \frac{1}{2} \left\langle \nabla^2 V(\mathbf{y}) \mathbf{h}, \mathbf{h} \right\rangle + \|\mathbf{h}\|^2 \cdot o(1) \tag{40}$$

uniformly in $\mathbf{y}$ on any compact set when $\mathbf{h} \to 0$. Assume for the moment that $c_V(\mathbf{x}, \mathbf{y}) \geq \frac{\delta}{2} \|\mathbf{x} - \mathbf{y}\|^2$ for some $\delta > 0$. Given any function $g \in \mathcal{C}^1$ with compact support and $\int g d\mu_{\tilde{V}} = 0$, introduce the infimal convolution associated with the cost c_V:

$$Q_c(g)(\mathbf{y}) := \inf_{\mathbf{x}} \{ g(\mathbf{x}) + c_V(\mathbf{x}, \mathbf{y}) \}, \tag{41}$$

whence $Q_c(g)(\mathbf{y}) - g(\mathbf{x}) \leq c_V(\mathbf{x}, \mathbf{y})$. By the definition of $\mathcal{W}_{c_V}$, for any joint probability measure π having marginals $\mu_{\tilde{V}}$ and ν, we must have

$$\mathcal{W}_{c_V}(\mu_{\tilde{V}}, \nu) = \inf_{\pi} \int c_V(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y}) \geq \int Q_c(g) d\nu - \int g d\mu_{\tilde{V}}. \tag{42}$$

Consider the infimum convolution of ϵg:

$$Q_c(\epsilon g)(\mathbf{y}) = \inf_{x} \{\epsilon g(\mathbf{x}) + c_V(\mathbf{x}, \mathbf{y})\} = \inf_{\mathbf{h}} \{\epsilon g(\mathbf{y} + \mathbf{h}) + c_V(\mathbf{y} + \mathbf{h}, \mathbf{y})\}. \tag{43}$$

Let $\mathbf{h}_\epsilon = \mathbf{h}_\epsilon(\mathbf{y})$ denote a point where the infimum is achieved. Since g is globally Lipschitz, say with constant L,

$$\epsilon g(\mathbf{y} + \mathbf{h}_\epsilon) + c_V(\mathbf{y} + \mathbf{h}_\epsilon, \mathbf{y}) \geq \epsilon g(\mathbf{y}) - \epsilon L \|\mathbf{h}_\epsilon\| + \delta \|\mathbf{h}_\epsilon\|^2. \tag{44}$$

On the other hand, by setting $\mathbf{h} = 0$ in (43), we see that $\epsilon g(\mathbf{y} + \mathbf{h}_\epsilon) + c_V(\mathbf{y} + \mathbf{h}_\epsilon, \mathbf{y}) \leq \epsilon g(\mathbf{y})$. Combining this with (44) gives

$$\|\mathbf{h}_\epsilon\| \leq \frac{L}{\delta} \epsilon. \tag{45}$$

Notice that (45) does not depend on y, and hence $\|\mathbf{h}_\epsilon\| = O(\epsilon)$ uniformly in $\mathbf{y}$.

As g is compactly supported, we have $\sup |g| = M < \infty$. Let Ω be the support of g, and let $B_\epsilon := \{\mathbf{x} \mid \exists \mathbf{y} \in \Omega, \|\mathbf{x} - \mathbf{y}\|^2 \leq \frac{2\epsilon M}{\delta}\}$. We claim that $Q_c(\epsilon g) \geq 0$ on B_ϵ^c. Indeed, for any $\mathbf{y} \in B_\epsilon^c$, suppose that the infimum of $\mathbf{x}$ in (43) is attained in Ω. Then $Q_c(\epsilon g) \geq \frac{\delta}{2} \cdot \frac{2\epsilon M}{\delta} - \epsilon M = 0$, since $c_V(\mathbf{x}, \mathbf{y}) \geq \frac{\delta}{2} \|\mathbf{x} - \mathbf{y}\|^2$ and $\inf_{\mathbf{x} \in \Omega} \|\mathbf{y} - \mathbf{x}\|^2 \geq \frac{2\epsilon M}{\delta}$. On the other hand, if the infimum of $\mathbf{x}$ in (43) is attained outside Ω, then $\epsilon g = 0$ and $c_V \geq 0$ implies that $Q_c(\epsilon g) \geq 0$.

For the sake of linearization, set $d\nu = (1 + \epsilon f)\, d\mu_{\tilde{V}}$ for some $f \in \mathcal{C}^1$ with $\int f d\mu_{\tilde{V}} = 0$. We now compute

$$\begin{aligned}
\mathcal{W}_{c_V}(\mu_{\tilde{V}}, (1 + \epsilon f) d\mu_{\tilde{V}}) &\geq \int Q_c(\epsilon g)\, (1 + \epsilon f)\, d\mu_{\tilde{V}} && \text{since } \int g d\mu_{\tilde{V}} = 0, \\
&\geq \int_{B_\epsilon} Q_c(\epsilon g)\, (1 + \epsilon f)\, d\mu_{\tilde{V}} && \text{since } Q_c(\epsilon g) \geq 0 \text{ on } B_\epsilon^c.
\end{aligned} \tag{46}$$

As the set B_ϵ is itself compact, we have, uniformly in $\mathbf{y}$,

$$\begin{aligned}
Q_c(\epsilon g)(\mathbf{y}) &= \epsilon g(\mathbf{y} + \mathbf{h}_\epsilon) + c(\mathbf{y} + \mathbf{h}_\epsilon, \mathbf{y}) \\
&= \epsilon g(\mathbf{y}) + \epsilon \left\langle \nabla g(\mathbf{y}), \mathbf{h}_\epsilon \right\rangle + \frac{1}{2} \left\langle \nabla^2 V(\mathbf{y}) \mathbf{h}_\epsilon, \mathbf{h}_\epsilon \right\rangle + o(\epsilon^2)
\end{aligned} \tag{47}$$

where the last line follows by (45). Noticing that (47) is convex in $\mathbf{h}_\epsilon$, we can find its minimum (up to $o(\epsilon^2)$) and write

$$Q_c(\epsilon g)(\mathbf{y}) \geq \epsilon g(\mathbf{y}) - \frac{\epsilon^2}{2} \left\langle \nabla^2 V^{-1}(\mathbf{y}) \nabla g(\mathbf{y}), \nabla g(\mathbf{y}) \right\rangle + o(\epsilon^2). \tag{48}$$

Multiplying (48) by $1 + \epsilon f$ and integrate on B_ϵ w.r.t. $d\mu_{\tilde{V}}$, we get, using (46) and $\int f d\mu_{\tilde{V}} = \int g d\mu_{\tilde{V}} = 0$,

$$\frac{1}{\epsilon^2} \mathcal{W}_{c_V}\left(\mu_{\tilde{V}}, (1 + \epsilon f)\, d\mu_{\tilde{V}}\right) \geq \int_{B_\epsilon} f g d\mu_{\tilde{V}} - \frac{1}{2} \int_{B_\epsilon} \left\langle \nabla^2 V^{-1} \nabla g, \nabla g \right\rangle d\mu_{\tilde{V}} + o(1). \tag{49}$$

By definition, B_ϵ contains the support of g, and hence the integrals in (49) are in fact over the whole space. We hence conclude

$$\liminf_{\epsilon\to 0} \frac{1}{\epsilon^2}\mathcal{W}_{c_V}\left(\mu_{\tilde{V}}, (1+\epsilon f)\, d\mu_{\tilde{V}}\right) \geq \int f g d\mu_{\tilde{V}} - \frac{1}{2}\int \langle \nabla^2 V^{-1} \nabla g, \nabla g\rangle \, d\mu_{\tilde{V}}. \tag{50}$$

Replacing g by λg in (50) and optimizing over λ, we get

$$\frac{\left(\int f g d\mu_{\tilde{V}}\right)^2}{2\int \langle \nabla V^{-1} \nabla g, \nabla g\rangle \, d\mu_{\tilde{V}}} \leq \liminf_{\epsilon\to 0} \frac{1}{\epsilon^2}\mathcal{W}_{c_V}\left(\mu_{\tilde{V}}, (1+\epsilon f)\, d\mu_{\tilde{V}}\right). \tag{51}$$

Moreover, using $\log(1+x) = x - \frac{x^2}{2} + o(x^2)$ and $\int f d\mu_{\tilde{V}} = 0$, we can compute

$$\begin{aligned} D\left((1+\epsilon f)\, d\mu_{\tilde{V}} \| \mu_{\tilde{V}}\right) &= \int (1+\epsilon f)\left(\epsilon f - \frac{\epsilon^2 f^2}{2} + o(\epsilon^2)\right) d\mu_{\tilde{V}} \\ &= \frac{\epsilon^2}{2}\int f^2 d\mu_{\tilde{V}} + o(\epsilon^2). \end{aligned} \tag{52}$$

In the case of $c_V(\mathbf{x},\mathbf{y}) \geq \frac{\delta}{2}\|\mathbf{x}-\mathbf{y}\|^2$, **Theorem 12** then follows by using $g = f$ in (51) and combing (39) and (52). For general case, replace V by $V + \frac{\delta\|\cdot\|^2}{2}$ and take $\delta \to 0$ in (50) and (52) to deduce the same inequalities.

■