This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3180298, IEEE Transactions on Information Theory

An Information-Theoretic View of Stochastic Localization

Ahmed El Alaoui and Andrea Montanari

***Abstract*—Given a probability measure μ over $\mathbb{R}^n$, it is often useful to approximate it by the convex combination of a small number of probability measures, such that each component is close to a product measure. Recently, Ronen Eldan used a stochastic localization argument to prove a general decomposition result of this type. In Eldan's theorem, the 'number of components' is characterized by the entropy of the mixture, and 'closeness to product' is characterized by the covariance matrix of each component.**

We present an elementary proof of Eldan's theorem which makes use of an information theory (or estimation theory) interpretation. The proof is analogous to the one of an earlier decomposition result known as the 'pinning lemma.'

***Index Terms*—Low complexity decompositions, the pinning lemma, stochastic localization.**

I. Motivation and result

LET μ be an arbitrary Borel probability measure on $\mathbb{R}^n$. A broadly useful approach to understanding μ is to decompose it into simpler components

$$\mu = \int_\Theta \mu_\theta \, \rho(\mathrm{d}\theta) =: \mathbb{E}_\theta \mu_\theta \, . \tag{I.1}$$

Here $(\Theta, \mathcal{F}_\Theta, \rho)$ is an abstract probability space, and, for each θ, μ_θ is a probability measure on $\mathbb{R}^n$. In this paper, the components μ_θ will be simple in the sense of being close to product measures.

Of course, the decomposition (I.1) is always possible: just take $\Theta = \mathbb{R}^n$, $\rho = \mu$, and $\mu_\theta = \delta_\theta$ (a point mass at θ) for $\theta \in \mathbb{R}^n$. This is a 'maximum entropy' decomposition (all the entropy of μ is pushed into ρ), and is not particularly useful. We will be instead interested in decompositions such that ρ has small entropy, and the μ_θ's are only approximately product measures.

Low-entropy decompositions have found applications in statistical mechanics [COP19], [Aus19], random graph theory [CD16], [Eld18], random constraint satisfaction problems [ACOGM20], high-dimensional statistics [COKPZ18], analysis of Markov chains [EKZ21] to name a few areas. A generic construction consists in letting μ_θ be the conditional distribution of $\boldsymbol{x} \sim \mu$ given a small subset of the coordinates of $\boldsymbol{x}$ (see below for further discussion). This leads to the so-called 'pinning lemma' which was discovered independently in [Mon08], [RT12]. Recently, Ronen Eldan [Eld20] pointed out that the pining lemma can be suboptimal and proved a general decomposition result with better properties. Eldan's proof follows the general approach of 'stochastic localization' which is in turn inspired by ideas in high-dimensional geometry [Eld16].

The purpose of this note is to present a new interpretation of Eldan's theorem, leading to an elementary proof. The interpretation has an information-theoretic (or estimation-theoretic) nature. We consider a noisy communication channel that takes as input $\boldsymbol{x} \sim \mu$ and outputs θ. The distribution μ_θ is then the conditional distribution of the channel input given its output θ. Note that this is the same interpretation as for the proof of the pinning lemma in [Mon08]. The main difference is that while in the pinning lemma the channel $\boldsymbol{x} \to \theta$ is an erasure channel, here we will use a Gaussian channel.

Turning to our construction, let $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ be a fixed positive semidefinite matrix, $\boldsymbol{x} \sim \mu$ and define the noisy observation (channel output) $\boldsymbol{y}$ via

$$\boldsymbol{y} = \sqrt{\tau}\, \boldsymbol{x} + \boldsymbol{Q}^{1/2} \boldsymbol{z} \, , \tag{I.2}$$

where $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_n)$, and τ is uniform in the interval $[1, 2]$. The random variables $\boldsymbol{x}$, $\boldsymbol{z}$ and τ are independent. For any fixed $\tau = t$, this is a noisy Gaussian channel, with signal-to-noise ratio (SNR) t. Introducing a random τ is convenient for the proof, but the estimates we prove hold also (with possibly worse constants) for all $\tau \in [1, 2]$, except on a set of small measure.

We set $\theta = (\boldsymbol{y}, \tau)$ and $\mu_\theta(\cdot) = \mu(\,\cdot\,|\theta)$. It is clear that $\mu = \mathbb{E}_\theta \mu_\theta$ so that the decomposition (I.1) holds.

Let ν be a background measure on $\mathbb{R}^n$ such that μ is absolutely continuous with respect to ν, i.e. $\mu \ll \nu$. Recall that the relative entropy of μ with respect to ν (Kullback-Leibler divergence) is defined by

$$\mathrm{D}(\mu \| \nu) = \mathbb{E}_\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\boldsymbol{x}). \tag{I.3}$$

It is easy to see that $\mu_\theta \ll \mu \ll \nu$ a.s., so that $\mathrm{D}(\mu_\theta \| \nu)$ is well defined.

Theorem 1 ([Eld20])**.** *We have*

$$\mathbb{E}_\theta \mathrm{Cov}(\mu_\theta) \preceq \boldsymbol{Q} \, , \tag{I.4}$$

$$0 \le \mathbb{E}_\theta \mathrm{D}(\mu_\theta \| \nu) - \mathrm{D}(\mu \| \nu) \tag{I.5}$$

$$\le \frac{1}{2} \log \det\big(\boldsymbol{I}_n + 2\boldsymbol{Q}^{-1} \mathrm{Cov}(\mu)\big) \, , \tag{I.6}$$

$$\mathbb{E}_\theta \big\{ \mathrm{Cov}(\mu_\theta) \boldsymbol{Q}^{-1} \mathrm{Cov}(\mu_\theta) \big\} \preceq \mathrm{Cov}(\mu) \, . \tag{I.7}$$

Remark I.1. Equations (I.4) and (I.7) bound the covariance of the component measure μ_θ. On the other hand, Eq. (I.6) controls the entropy of the variable θ, which is given by the difference between the entropy of the mixture and the

A. El Alaoui is in the Department of Statistics and Data Science, Cornell University, Ithaca, NY 14850, USA. Email: elalaoui@cornell.edu.

A. Montanari is in the Department of Electrical Engineering and the Department of Statistics, Stanford University, CA 94305, USA. Email: montanar@stanford.edu.

Manuscript received September 8, 2021; revised June 1, 2022.

0018-9448 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Stanford University. Downloaded on July 26,2022 at 04:39:17 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3180298, IEEE Transactions on Information Theory

average entropy of the components. More precisely, in information theoretic terms, we control the mutual information $\mathbb{E}_\theta \mathrm{D}(\mu_\theta\|\nu) - \mathrm{D}(\mu\|\nu) = \mathrm{I}(\theta;\boldsymbol{x})$ (see proof for definitions).

Remark I.2. As mentioned above, the difference between the decomposition presented here and the pinning lemma of [Mon08] is in the choice of the noisy channel. Here we use the Gaussian channel (I.2) (neglecting for a moment the fact that τ is random). In contrast, in [Mon08], the channel is an erasure channel: $y_i = x_i$ with probability ε and $y_i = *$ (an erasure) otherwise, independently across coordinates.

Remark I.3. The expectation with respect to θ in Eqs. (I.4) to (I.7) is an expectation with respect to $(\tau, \boldsymbol{y})$: $\mathbb{E}_\theta(\,\cdot\,) = \mathbb{E}_\tau(\mathbb{E}_{\boldsymbol{y}|\tau}(\,\cdot\,|\tau))$. The expectation with respect to τ can be eliminated using Markov inequality. This implies that there exists a set $T \subseteq [1,2]$ of Lebesgue measure $|T| \ge 1 - 3M^{-1}$ such that, for any $t \in T$, the inequalities (I.4), (I.6), (I.7) hold up to a factor M for $\tau = t$. For instance, the first inequality is replaced by $\mathbb{E}(\mathrm{Cov}(\mu_\theta)|\tau = t) \preceq M\boldsymbol{Q}$. Note that the conditional expectation is simply the expectation with respect to $\boldsymbol{y} = \sqrt{t}\boldsymbol{x} + \boldsymbol{Q}^{1/2}\boldsymbol{z}$.

We finally notice that the above information-theoretic interpretation does not only apply to the final decomposition (I.1), but to the whole measure-valued stochastic process defined in [Eld20] [1]. We explain this extension in Section III.

II. Proof

A. Proof of Eq. (I.4)

We compare the error of the maximum likelihood estimator of $\boldsymbol{x}$ to that of the Bayes optimal estimator of $\boldsymbol{x}$ given $\theta = (\boldsymbol{y}, \tau)$:

$$\widehat{\boldsymbol{x}}_{\mathrm{ML}} = \tau^{-1/2}\,\boldsymbol{y} \quad \text{and} \quad \widehat{\boldsymbol{x}}_{\mathrm{Bayes}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y},\tau].$$

Let $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ be a fixed PSD matrix, and define $\|\boldsymbol{v}\|_{\boldsymbol{R}}^2 = \langle \boldsymbol{v}, \boldsymbol{R}\boldsymbol{v}\rangle$. Since $\widehat{\boldsymbol{x}}_{\mathrm{Bayes}}$ minimizes the mean squared error $\mathbb{E}\{\|\boldsymbol{x} - \widehat{\boldsymbol{x}}(\theta)\|_{\boldsymbol{R}}^2\}$ among all measurable estimators $\widehat{\boldsymbol{x}}$, we have

$$\mathbb{E}\mathrm{Tr}\big[\boldsymbol{R}(\boldsymbol{x} - \widehat{\boldsymbol{x}}_{\mathrm{Bayes}})(\boldsymbol{x} - \widehat{\boldsymbol{x}}_{\mathrm{Bayes}})^\top\big] \le \mathbb{E}\mathrm{Tr}\big[\boldsymbol{R}(\boldsymbol{x} - \widehat{\boldsymbol{x}}_{\mathrm{ML}})(\boldsymbol{x} - \widehat{\boldsymbol{x}}_{\mathrm{ML}})^\top\big]\,. \quad \text{(II.1)}$$

The left-hand side is equal to $\mathbb{E}\mathrm{Tr}(\boldsymbol{R}\mathrm{Cov}(\mu_\theta))$, and the right-hand side is equal to

$$\mathbb{E}\big[\tau^{-1}\langle \boldsymbol{z}\boldsymbol{Q}^{1/2}, \boldsymbol{R}\boldsymbol{Q}^{1/2}\boldsymbol{z}\rangle\big] = \mathbb{E}[\tau^{-1}]\cdot\mathrm{Tr}(\boldsymbol{R}\boldsymbol{Q}). \quad \text{(II.2)}$$

Since the inequality holds for all $\boldsymbol{R} \succeq 0$, we conclude that

$$\mathbb{E}\,\mathrm{Cov}(\mu_\theta) \preceq \mathbb{E}[\tau^{-1}]\cdot\boldsymbol{Q} \preceq \boldsymbol{Q}\,. \quad \text{(II.3)}$$

B. Proof of (I.5) and (I.6)

This claim follows immediately using some basic inequalities from information theory [CT91], [DCT91].

[1] After the completion and submission of this work, we learned that this extension has been independently discovered in a concurrent work of Klartag and Putterman [KP21].

The mutual information of two random variables X, Y is $\mathrm{I}(X;Y) := \mathrm{D}(\mu_{X,Y}\|\mu_X \times \mu_Y)$, where we denote by $\mu_{X,Y,\dots}$ the joint law of $X, Y, \dots$. If $\mu_X, \mu_{X|Y}(\,\cdot\,|y) \ll \nu_X$, we have

$$\mathrm{I}(X;Y) := \mathbb{E}_y \mathrm{D}(\mu_{X|Y}(\,\cdot\,|y)\|\nu_X) - \mathrm{D}(\mu_X\|\nu_X)\,. \quad \text{(II.4)}$$

We consider $X = \boldsymbol{x}$, $Y = \boldsymbol{y}$, under their joint distribution given $\tau = t$. In other words $\boldsymbol{x} \sim \mu$ and $\boldsymbol{y}$ is given by Eq. (I.2) with $\tau = t$. We have $\mu_\theta = \mu_{X|Y}(\,\cdot\,|\boldsymbol{y})$. Using non-negativity of the mutual information, we have

$$0 \le \mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{y}}\mathrm{D}(\mu_\theta\|\nu) - \mathrm{D}(\mu\|\nu)\,. \quad \text{(II.5)}$$

which yields the inequality Eq. (I.5). As for Eq. (I.6), inverting the role of X, Y in the definition of mutual information, and letting $\overline{\nu}$ be, for instance, the standard Gaussian measure we get

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}}\mathrm{D}(\mu_{\boldsymbol{y}|\boldsymbol{x}}(\,\cdot\,|\boldsymbol{x})\|\overline{\nu}) - \mathrm{D}(\mu_{\boldsymbol{y}}\|\overline{\nu})\,. \quad \text{(II.6)}$$

Recall the definition of differential entropy of a random variable X with density f with respect to the Lebesgue measure: $h(X) := -\int f(x)\log f(x)\mathrm{d}x$, and $h(X|Y) = h(X,Y) - h(Y)$. We then have from the last display

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = h(\boldsymbol{y}) - h(\boldsymbol{y}|\boldsymbol{x})\,. \quad \text{(II.7)}$$

The differential entropy of an n-dimensional Gaussian vector $\boldsymbol{g}$ with covariance $\boldsymbol{\Sigma}$ is $h(\boldsymbol{g}) = (n/2)\log(2\pi e) + (1/2)\mathrm{Tr}\log\boldsymbol{\Sigma}$. Therefore,

$$h(\boldsymbol{y}|\boldsymbol{x}) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\mathrm{Tr}\log\boldsymbol{Q}\,.$$

Moreover, the Gaussian distribution maximizes the differential entropy among all those distributions with the same covariance, hence

$$h(\boldsymbol{y}) \le \frac{n}{2}\log(2\pi e) + \frac{1}{2}\mathrm{Tr}\log\mathrm{Cov}(\mu_{\boldsymbol{y}})\,. \quad \text{(II.8)}$$

Since $\mathrm{Cov}(\mu_{\boldsymbol{y}}) = t\mathrm{Cov}(\mu) + \boldsymbol{Q}$, we obtain

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \le \frac{1}{2}\mathrm{Tr}\log\mathrm{Cov}(\mu_{\boldsymbol{y}}) - \frac{1}{2}\mathrm{Tr}\log\boldsymbol{Q} \quad \text{(II.9)}$$

$$= \frac{1}{2}\log\det\big(\boldsymbol{I}_n + t\boldsymbol{Q}^{-1}\mathrm{Cov}(\mu)\big)\,. \quad \text{(II.10)}$$

The claim then follows since $t \le 2$.

C. Proof of (I.7)

Fixing $\tau = t$ a non-random value, we write $\boldsymbol{y}_t$ for the corresponding output of channel (I.2), namely $\boldsymbol{y}_t = \sqrt{t}\boldsymbol{x} + \boldsymbol{Q}^{1/2}\boldsymbol{z}$. We denote by $\mu_{\boldsymbol{y}_t}(\mathrm{d}\boldsymbol{x})$ for the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}_t$. In order to emphasize its dependence on t, we will write $\mu_{\boldsymbol{x}_0,\boldsymbol{z},t} := \mu_{\boldsymbol{y}_t = \sqrt{t}\boldsymbol{x}_0 + \boldsymbol{Q}^{1/2}\boldsymbol{z}}$. Simplifying Bayes formula, we get

$$\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\mathrm{d}\boldsymbol{x}) = \exp\Big\{-\frac{t}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_{\boldsymbol{Q}^{-1}}^2 + \sqrt{t}\langle \boldsymbol{z}, \boldsymbol{Q}^{-1/2}\boldsymbol{x}\rangle\Big\}\frac{\mu(\mathrm{d}\boldsymbol{x})}{Z_{\boldsymbol{x}_0,\boldsymbol{z},t}}\,. \quad \text{(II.11)}$$

Here we use the notation $\|\boldsymbol{v}\|_{\boldsymbol{A}}^2 := \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}\rangle$.

0018-9448 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Stanford University. Downloaded on July 26,2022 at 04:39:17 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3180298, IEEE Transactions on Information Theory

Throughout this proof, given a measure $\nu(\mathrm{d}\boldsymbol{x})$ and function $\psi_1(\boldsymbol{x}), \psi_2(\boldsymbol{x})$, we use the shorthands

$$\nu(\psi(\boldsymbol{x})) := \int \psi(\boldsymbol{x})\nu(\mathrm{d}\boldsymbol{x}) \quad \text{and,}$$
$$\nu(\psi_1(\boldsymbol{x}); \psi_2(\boldsymbol{x})) := \nu(\psi_1(\boldsymbol{x}) \cdot \psi_2(\boldsymbol{x})) - \nu(\psi_1(\boldsymbol{x}))\nu(\psi_2(\boldsymbol{x}))\,.$$

We then have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x}) = -\frac{1}{2}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}\big(\boldsymbol{x}; \|\boldsymbol{x}-\boldsymbol{x}_0\|^2_{\boldsymbol{Q}^{-1}}\big) + \frac{1}{2\sqrt{t}}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}\big(\boldsymbol{x}; \langle \boldsymbol{z}, \boldsymbol{Q}^{-1/2}\boldsymbol{x}\rangle\big)\,. \quad \text{(II.12)}$$

(Note that $\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x})$ is the mean of the vector $\boldsymbol{x}$.)

Given a matrix $\boldsymbol{R} \succeq 0$, we define the minimum mean square error

$$\begin{aligned}\mathrm{mmse}(t) &:= \mathbb{E}\Big\{\big\|\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}|\boldsymbol{y}_t)\big\|^2_{\boldsymbol{R}}\Big\} \quad \text{(II.13)}\\ &= \mathrm{Tr}\big(\boldsymbol{R}\,\mathrm{Cov}(\mu)\big) - \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z}}\langle \mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x}), \boldsymbol{R}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x})\rangle\,.\end{aligned}$$

Differentiating, and using the formula above to differentiate $\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x})$, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{mmse}(t) &= -2\cdot \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z}}\langle \mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x}), \boldsymbol{R}\frac{\mathrm{d}}{\mathrm{d}t}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x})\rangle\\ &=: A - B_1 + B_2\,, \quad \text{(II.14)}\end{aligned}$$

where

$$\begin{aligned}A &:= \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z}}\big\{\langle \mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x}), \boldsymbol{R}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x}; \|\boldsymbol{x}-\boldsymbol{x}_0\|^2_{\boldsymbol{Q}^{-1}})\rangle\big\}\,,\\ B_1 &:= \frac{1}{\sqrt{t}}\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z}}\big\{\langle \mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x}), \boldsymbol{R}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x}\langle \boldsymbol{z}, \boldsymbol{Q}^{-1/2}\boldsymbol{x}\rangle)\rangle\big\}\,,\\ B_2 &:= \frac{1}{\sqrt{t}}\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z}}\big\{\langle \mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x}), \boldsymbol{R}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\boldsymbol{x})\rangle \cdot \mu_{\boldsymbol{x}_0,\boldsymbol{z},t}(\langle \boldsymbol{z}, \boldsymbol{Q}^{-1/2}\boldsymbol{x}\rangle)\big\}\,.\end{aligned}$$

We next use Gaussian integration by parts (Stein's Lemma) to simplify the terms B_1, B_2. It is useful to note that

$$\frac{1}{\sqrt{t}}\nabla_{\boldsymbol{z}}\mu_{\boldsymbol{x}_0,\boldsymbol{z},t}\big(\psi(\boldsymbol{x})\big) = \mu_{\boldsymbol{x}_0,\boldsymbol{z},t}\big(\psi(\boldsymbol{x}); \boldsymbol{Q}^{-1/2}\boldsymbol{x}\big)\,. \quad \text{(II.15)}$$

Writing for simplicity $\mu_t := \mu_{\boldsymbol{x}_0,\boldsymbol{z},t}$ and $\mathbb{E} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z}}$, we get

$$\begin{aligned}B_1 &= \mathbb{E}\big\{\mu_t(\boldsymbol{x})^\top \boldsymbol{R}\mu_t(\boldsymbol{x}\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}})\big\} \quad \text{(II.16)}\\ &\quad - \mathbb{E}\big\{\mu_t(\boldsymbol{x})^\top \boldsymbol{R}\mu_t(\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{Q}^{-1}\mu_t(\boldsymbol{x})\big\}\\ &\quad + \mathbb{E}\mathrm{Tr}\big\{\boldsymbol{Q}^{-1}\big[\mu_t(\boldsymbol{x}\boldsymbol{x}^\top) - \mu_t(\boldsymbol{x})\mu_t(\boldsymbol{x}^\top)\big]\boldsymbol{R}\mu_t(\boldsymbol{x}\boldsymbol{x}^\top)\big\}\,,\\ B_2 &= \mathbb{E}\big\{\mu_t(\boldsymbol{x})^\top \boldsymbol{R}\mu_t(\boldsymbol{x})\mu_t(\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}})\big\} \quad \text{(II.17)}\\ &\quad - \mathbb{E}\big\{\mu_t(\boldsymbol{x})^\top \boldsymbol{R}\mu_t(\boldsymbol{x})\mu_t(\boldsymbol{x})^\top\boldsymbol{Q}^{-1}\mu_t(\boldsymbol{x})\big\}\\ &\quad + 2\mathbb{E}\big\{\mu_t(\boldsymbol{x})^\mathsf{T}\boldsymbol{Q}^{-1}[\mu_t(\boldsymbol{x}\boldsymbol{x}^\top) - \mu_t(\boldsymbol{x})\mu_t(\boldsymbol{x})^\top]\boldsymbol{R}\mu_t(\boldsymbol{x})\big\}\,.\end{aligned}$$

Finally, by the tower property of conditional expectation

$$\begin{aligned}A &= \mathbb{E}\big\{\mu_t(\boldsymbol{x})^\top \boldsymbol{R}\mu_t(\boldsymbol{x}; \|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}})\big\}\\ &\quad - 2\mathbb{E}\big\{\mu_t(\boldsymbol{x})^\top \boldsymbol{R}\mu_t(\boldsymbol{x}; \boldsymbol{x}^\top)\boldsymbol{Q}^{-1}\boldsymbol{x}_0\big\}\\ &= \mathbb{E}\big\{\mu_t(\boldsymbol{x})^\top \boldsymbol{R}\mu_t(\boldsymbol{x}; \|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}})\big\} \quad \text{(II.18)}\\ &\quad - 2\mathbb{E}\big\{\mu_t(\boldsymbol{x})^\top \boldsymbol{R}[\mu_t(\boldsymbol{x}\boldsymbol{x}^\top) - \mu_t(\boldsymbol{x})\mu_t(\boldsymbol{x})^\top]\boldsymbol{Q}^{-1}\mu_t(\boldsymbol{x})\big\}\,.\end{aligned}$$

We next substitute Eqs. (II.16), (II.17), and (II.18) in Eq. (II.14) to get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{mmse}(t) = -\mathbb{E}\mathrm{Tr}\big\{\mathrm{Cov}(\mu_t)\boldsymbol{Q}^{-1}\mathrm{Cov}(\mu_t)\boldsymbol{R}\big\}\,. \quad \text{(II.19)}$$

We next integrate this identity for $t \in [1,2]$, to get

$$\begin{aligned}\int_1^2 \mathbb{E}\mathrm{Tr}\big\{\mathrm{Cov}(\mu_t)\boldsymbol{Q}^{-1}\mathrm{Cov}(\mu_t)\boldsymbol{R}\big\}\mathrm{d}t &= \mathrm{mmse}(1) - \mathrm{mmse}(2)\\ &\le \mathrm{mmse}(0) = \mathrm{Tr}(\mathrm{Cov}(\mu)\boldsymbol{R})\,.\end{aligned}$$

Finally, the integral over t can be interpreted as an expectation over $\tau \sim \mathsf{Unif}([1,2])$, whence

$$\mathbb{E}\mathrm{Tr}\big\{\mathrm{Cov}(\mu_\theta)\boldsymbol{Q}^{-1}\mathrm{Cov}(\mu_\theta)\boldsymbol{R}\big\} \le \mathrm{Tr}(\mathrm{Cov}(\mu)\boldsymbol{R})\,.$$

Since this holds for any $\boldsymbol{R} \succeq 0$, we proved Eq. (I.7).

III. Extension to the stochastic localization process

Another way of writing the output of the Gaussian channel is as follows: Let $\boldsymbol{x} \sim \mu$ and $(\boldsymbol{B}_t)_{t\ge 0}$ be a standard Brownian motion in $\mathbb{R}^n$. Further, let

$$\bar{\boldsymbol{y}}_t = t\boldsymbol{x} + \boldsymbol{Q}^{1/2}\boldsymbol{B}_t. \quad \text{(III.1)}$$

Then it is clear that $\bar{\boldsymbol{y}}_t$ has the same law as the vector $\sqrt{t}\boldsymbol{y}$, so the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ has the same law as the probability measure

$$\mu_t(\mathrm{d}\boldsymbol{x}) = \frac{1}{Z_t}\exp\Big\{\langle \bar{\boldsymbol{y}}_t, \boldsymbol{x}\rangle_{\boldsymbol{Q}^{-1}} - \frac{t}{2}\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}}\Big\}\mu(\mathrm{d}\boldsymbol{x})\,. \quad \text{(III.2)}$$

We will show that the measure-valued process $(\mu_t)_{t\ge 0}$ satisfies the stochastic localization SDE of Eldan [Eld20], as stated below.

Theorem 2. *Write L_t for the likelihood ratio process of μ_t with respect to μ:*

$$L_t(\boldsymbol{x}) := \frac{\mathrm{d}\mu_t}{\mathrm{d}\mu}(\boldsymbol{x})\,. \quad \text{(III.3)}$$

Then there exist a Brownian motion $(\boldsymbol{W}_t)_{t\ge 0}$ adapted to the filtration generated by $(\bar{\boldsymbol{y}}_t)_{t\ge 0}$, such that for all $\boldsymbol{x} \in \mathbb{R}^n$ and $t \ge 0$, we have

$$\mathrm{d}L_t(\boldsymbol{x}) = L_t(\boldsymbol{x})\big\langle \boldsymbol{x} - \boldsymbol{a}_t, \boldsymbol{Q}^{-1/2}\mathrm{d}\boldsymbol{W}_t\big\rangle\,, \quad \textit{and} \quad L_0(\boldsymbol{x}) = 1\,, \quad \text{(III.4)}$$

where

$$\boldsymbol{a}_t = \mathbb{E}\big\{\boldsymbol{x}|\bar{\boldsymbol{y}}_t\big\} = \int \boldsymbol{x}\,\mu_t(\mathrm{d}\boldsymbol{x})\,. \quad \text{(III.5)}$$

Proof. We use the representation (III.2) to write

$$\mathrm{d}\log L_t(\boldsymbol{x}) = \langle \mathrm{d}\bar{\boldsymbol{y}}_t, \boldsymbol{x}\rangle_{\boldsymbol{Q}^{-1}} - \frac{1}{2}\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}}\mathrm{d}t - \mathrm{d}\log Z_t\,. \quad \text{(III.6)}$$

Let us write $h_t(\boldsymbol{x})$ for the expression appearing in the exponent in Eq. (III.2). By Itô's formula we have

$$\begin{aligned}\mathrm{d}Z_t &= \int_{\mathbb{R}^n}\Big(\langle \boldsymbol{x}, \boldsymbol{Q}^{-1}\mathrm{d}\bar{\boldsymbol{y}}_t\rangle - \frac{1}{2}\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}}\mathrm{d}t\Big)e^{h_t(\boldsymbol{x})}\mu(\mathrm{d}\boldsymbol{x})\\ &\quad + \frac{1}{2}\Big(\int_{\mathbb{R}^n}\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}}e^{h_t(\boldsymbol{x})}\mu(\mathrm{d}\boldsymbol{x})\Big)\mathrm{d}t\\ &= \Big\langle \int_{\mathbb{R}^n}\boldsymbol{x}e^{h_t(\boldsymbol{x})}\mu(\mathrm{d}\boldsymbol{x}), \boldsymbol{Q}^{-1}\mathrm{d}\bar{\boldsymbol{y}}_t\Big\rangle\,. \quad \text{(III.7)}\end{aligned}$$

0018-9448 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Stanford University. Downloaded on July 26,2022 at 04:39:17 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3180298, IEEE Transactions on Information Theory

With the notation $\boldsymbol{a}_t = \mathbb{E}\{\boldsymbol{x}|\bar{\boldsymbol{y}}_t\} = \int \boldsymbol{x}\, \mu_t(\mathrm{d}\boldsymbol{x})$ (and writing $[Z]_t$ for the quadratic variation process), we obtain

$$\mathrm{d}\log Z_t = \frac{\mathrm{d}Z_t}{Z_t} - \frac{1}{2}\frac{\mathrm{d}[Z]_t}{Z_t^2} \tag{III.8}$$

$$= \big\langle \boldsymbol{a}_t, \mathrm{d}\bar{\boldsymbol{y}}_t \big\rangle_{\boldsymbol{Q}^{-1}} - \frac{1}{2}\|\boldsymbol{a}_t\|^2_{\boldsymbol{Q}^{-1}}\mathrm{d}t\,. \tag{III.9}$$

Substituting this in Eq. (III.6) we have

$$\begin{aligned}\mathrm{d}\log L_t(\boldsymbol{x}) &= \big\langle \boldsymbol{x}-\boldsymbol{a}_t, \mathrm{d}\bar{\boldsymbol{y}}_t\big\rangle_{\boldsymbol{Q}^{-1}} - \frac{1}{2}\Big(\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}} - \|\boldsymbol{a}_t\|^2_{\boldsymbol{Q}^{-1}}\Big)\mathrm{d}t \\ &= \big\langle \boldsymbol{x}-\boldsymbol{a}_t, \mathrm{d}\bar{\boldsymbol{y}}_t - \boldsymbol{a}_t\mathrm{d}t\big\rangle_{\boldsymbol{Q}^{-1}} - \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{a}_t\|^2_{\boldsymbol{Q}^{-1}}\mathrm{d}t\,.\end{aligned} \tag{III.10}$$

It remains to understand the law of the process $\bar{\boldsymbol{y}}_t - \int_0^t \boldsymbol{a}_s \mathrm{d}s$. We let $\mathcal{F}_t = \sigma(\{\bar{\boldsymbol{y}}_s : 0 \le s \le t\})$ and $\mathbb{F} = (\mathcal{F}_t)_{t\ge 0}$. It is known that the process $(\boldsymbol{W}_t)_{t\ge 0}$ defined by

$$\boldsymbol{W}_t := \boldsymbol{Q}^{-1/2}\Big(\bar{\boldsymbol{y}}_t - \int_0^t \mathbb{E}\{\boldsymbol{x}|\mathcal{F}_s\}\mathrm{d}s\Big)\,, \tag{III.11}$$

is an $\mathbb{F}$-adapted Brownian motion. See for instance Theorem 7.12 in [LS77], or Theorem 5.13 in [LG16]. We conclude by arguing that $\mathbb{E}\{\boldsymbol{x}|\mathcal{F}_t\} = \boldsymbol{a}_t$.

Indeed, for a fixed $t \ge 0$, we let

$$p_{t,\boldsymbol{x}} := \mathcal{L}\big((\bar{\boldsymbol{y}}_s)_{s\in[0,t]} \,\big|\, \boldsymbol{x}\big)$$

denote the law of the path $(\bar{\boldsymbol{y}}_s)_{s\in[0,T]}$ conditional on $\boldsymbol{x}$ as per Eq. (III.1), and let $q_t = p_{t,\boldsymbol{x}=0}$. We obtain by Girsanov's theorem—see also Theorem 7.1 and its Corollary in [LS77]—that

$$\frac{\mathrm{d}p_{t,\boldsymbol{x}}}{\mathrm{d}q_t}(\boldsymbol{y}) = \exp\Big\{\langle \boldsymbol{y}_t, \boldsymbol{x}\rangle_{\boldsymbol{Q}^{-1}} - \frac{t}{2}\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^{-1}}\Big\}\,, \tag{III.12}$$

for all $\boldsymbol{y} \in C([0,t],\mathbb{R}^n)$ (the space of continuous functions on $[0,t]$ endowed with the topology of uniform convergence). We observe that the above only depends on the path $\boldsymbol{y}$ through the endpoint $\boldsymbol{y}_t$. Therefore by the Bayes rule, the conditional distribution of $\boldsymbol{x}$ given $\mathcal{F}_t$ is equal to μ_t (III.2), and we have $\mathbb{E}\{\boldsymbol{x}|\mathcal{F}_t\} = \mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}_t\} = \boldsymbol{a}_t$ as a consequence. Continuing from Eq. (III.10), we have

$$\mathrm{d}\log L_t(\boldsymbol{x}) = \big\langle \boldsymbol{x}-\boldsymbol{a}_t, \boldsymbol{Q}^{-1/2}\mathrm{d}\boldsymbol{W}_t\big\rangle - \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{a}_t\|^2_{\boldsymbol{Q}^{-1}}\mathrm{d}t\,, \tag{III.13}$$

and we obtain (III.4) by applying Itô's formula once more. □

ACKNOWLEDGMENT

We thank Ronen Eldan for a stimulating discussion. This work was carried out while the authors were (virtually) participating in the Fall 2020 program on Probability, Geometry and Computation in High Dimensions at the Simons Institute for the Theory of Computing. A.EA. was supported by the Simons Berkeley Richard M. Karp Research Fellowship. A.M. was partially supported by the NSF grant CCF-2006489 and the ONR grant N00014-18-1-2729.

REFERENCES

[ACOGM20] Peter Ayre, Amin Coja-Oghlan, Pu Gao, and Noëla Müller, *The satisfiability threshold for random linear equations*, Combinatorica **40** (2020), no. 2, 179–235.

[Aus19] Tim Austin, *The structure of low-complexity gibbs measures on product spaces*, The Annals of Probability **47** (2019), no. 6, 4002–4023.

[CD16] Sourav Chatterjee and Amir Dembo, *Nonlinear large deviations*, Advances in Mathematics **299** (2016), 396–450.

[COKPZ18] Amin Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborová, *Information-theoretic thresholds from the cavity method*, Advances in Mathematics **333** (2018), 694–795.

[COP19] Amin Coja-Oghlan and Will Perkins, *Bethe states of random factor graphs*, Communications in Mathematical Physics **366** (2019), no. 1, 173–201.

[CT91] Thomas. M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[DCT91] Amir Dembo, Thomas M Cover, and Joy A Thomas, *Information theoretic inequalities*, IEEE Transactions on Information theory **37** (1991), no. 6, 1501–1518.

[EKZ21] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni, *A spectral condition for spectral gap: fast mixing in high-temperature ising models*, Probability Theory and Related Fields (2021), 1–17.

[Eld16] Ronen Eldan, *Skorokhod embeddings via stochastic flows on the space of gaussian measures*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, vol. 52, Institut Henri Poincaré, 2016, pp. 1259–1280.

[Eld18] ———, *Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations*, Geometric and Functional Analysis **28** (2018), no. 6, 1548–1596.

[Eld20] ———, *Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation*, Probability Theory and Related Fields **176** (2020), no. 3, 737–755.

[KP21] Bo'az Klartag and Eli Putterman, *Spectral monotonicity under Gaussian convolution*, arXiv preprint arXiv:2107.09496, 2021.

[LG16] Jean-François Le Gall, *Brownian motion, martingales, and stochastic calculus*, Springer, 2016.

[LS77] Robert Shevilevich Liptser and Al'bert Nikolaevich Shiriaev, *Statistics of random processes: General theory*, vol. 394, Springer, 1977.

[Mon08] Andrea Montanari, *Estimating random variables from random sparse observations*, European Transactions on Telecommunications **19** (2008), 385–403.

[RT12] Prasad Raghavendra and Ning Tan, *Approximating csps with global cardinality constraints using sdp hierarchies*, Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms, SIAM, 2012, pp. 373–387.

0018-9448 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Stanford University. Downloaded on July 26,2022 at 04:39:17 UTC from IEEE Xplore. Restrictions apply.