6.S891 Lecture 25: Nonlinear Large Deviations

Kuikui Liu

December 12, 2023

In this lecture, we return to variational approximations of mean-field type. The focus here is on applications to *large deviations* problems. One should have in mind some "simple" probability measure ν on some "simple" state space Ω (e.g. $\mathsf{Unif}\{\pm 1\}^n$), and some concrete function f of interest which is *nonlinear*. The goal is to control the upper tail probability $\Pr_{x\sim\nu}[f(x) \geq (1+\delta)\mathbb{E}_\nu[f]]$ to leading order for constant $\delta > 0$. We previously saw how to bound this assuming there is a reversible local Markov chain w.r.t. ν which satisfies a modified log-Sobolev inequality. This method is "robust" in the sense that it holds simultaneously for all 1-Lipschitz functions, but in general, it does not give the sharpest bounds for specific test functions encountered in applications. Here, we use variational principles to go beyond these limitations.

1 Revisiting the Naïve Mean-Field Approximation

For convenience, we again consider probability measures over the family $2^{\mathcal{U}}$ of all subsets of some ground set $\mathcal{U}$ (e.g. $[n]$). We endow this space with some "background" probability measure ν (e.g. $\mathsf{Unif}2^{\mathcal{U}}$), which we leave unspecified for a moment. For a function $f : 2^{\mathcal{U}} \to \mathbb{R}$ (the "Hamiltonian"), recall that the Gibbs Variational Principle states that

$$\mathcal{F}_\nu(f) \overset{\mathsf{def}}{=} \log \mathbb{E}_\nu\left[e^f\right] = \sup_\zeta \left\{\mathbb{E}_\zeta[f] - \mathscr{D}_{\mathrm{KL}}\left(\zeta \,\|\, \nu\right)\right\}, \tag{1}$$

where the supremum over all probability measures ζ over $2^{\mathcal{U}}$. Since $2^{\mathcal{U}}$ is a product space, if ν is a product measure (e.g. the *p-biased* measure $\nu_p(S) \overset{\mathsf{def}}{=} p^{|S|}(1-p)^{|\mathcal{U}\setminus S|}$ for $p \in [0,1]$), then it makes sense to restrict the above convex program to product measures, i.e.

$$\mathcal{F}_\nu^{\mathsf{NMF}}(f) \overset{\mathsf{def}}{=} \sup_{\boldsymbol{m}\in[0,1]^{\mathcal{U}}} \left\{\mathbb{E}_{\pi(\boldsymbol{m})}[f] - \mathscr{D}_{\mathrm{KL}}\left(\pi(\boldsymbol{m}) \,\|\, \nu\right)\right\}, \tag{2}$$

where $\pi(\boldsymbol{m})$ denotes the unique product measure over $2^{\mathcal{U}}$ with coordinate marginals $\boldsymbol{m}$ (i.e. $\Pr_{S\sim\pi(\boldsymbol{m})}[i \in S] = \boldsymbol{m}_i$). Recall that we say the naïve mean-field approximation is "correct" if $\frac{\mathcal{F}_\nu(f) - \mathcal{F}_\nu^{\mathsf{NMF}}(f)}{|\mathcal{U}|} \leq o(1)$.

The relevance of this to large deviations is via the standard connection between concentration phenomena and bounds on the moment generating function.

Lemma 1.1 (Informal)**.** *Assume* $\frac{\mathcal{F}_\nu(s\cdot f) - \mathcal{F}_\nu^{\mathsf{NMF}}(s\cdot f)}{|\mathcal{U}|} \leq o(1)$ *for all* $s > 0$*. Then*

$$\log \Pr_{S\sim\nu}[f(S) \geq t] \leq -\inf_{\boldsymbol{m}\in[0,1]^{\mathcal{U}}} \left\{\mathscr{D}_{\mathrm{KL}}\left(\pi(\boldsymbol{m}) \,\|\, \nu\right) \,\middle|\, \mathbb{E}_{\pi(\boldsymbol{m})}[f] \geq t\right\} + o\left(|\mathcal{U}|\right). \tag{3}$$

Note that $\mathscr{D}_{\mathrm{KL}}\left(\pi(\boldsymbol{m}) \,\|\, \nu\right) = \sum_{i\in\mathcal{U}} \left(\boldsymbol{m}_i \log \frac{\boldsymbol{m}_i}{\nu_i} + (1-\boldsymbol{m}_i)\log\frac{1-\boldsymbol{m}_i}{1-\nu_i}\right)$ *since both arguments are product measures.*

Proof. For every fixed parameter $s > 0$, we have

$$\begin{aligned}\log \Pr_{S\sim\nu}[f(S) \geq t] &= \log \Pr_{S\sim\nu}\left[e^{s\cdot f(S)} \geq e^{s\cdot t}\right] \\ &\leq \log \mathbb{E}_\nu\left[e^{s\cdot f}\right] - s\cdot t. && \text{(Markov's Inequality)}\end{aligned}$$

Selecting the best choice of s and applying the Gibbs Variational Principle Eq. (1), we have

$$\begin{aligned}\log \Pr_{S\sim\nu}[f(S)\geq t] &\leq \inf_{s>0}\sup_{\zeta}\{s\cdot\mathbb{E}_\zeta[f]-\mathscr{D}_{\mathrm{KL}}(\zeta\,\|\,\nu)-s\cdot t\}\\ &= \sup_{\zeta}\left\{-\mathscr{D}_{\mathrm{KL}}(\zeta\,\|\,\nu)+\inf_{s>0}\{s\cdot(\mathbb{E}_\zeta[f]-t)\}\right\}\\ &\qquad\qquad\text{(von Neumann's Minimax Theorem)}\\ &= -\inf_{\zeta}\left\{\mathscr{D}_{\mathrm{KL}}(\zeta\,\|\,\nu)\;\middle|\;\mathbb{E}_\zeta[f]\geq t\right\},\end{aligned}$$

where in the final step, we used the fact that if $\mathbb{E}_\zeta[f]<t$, then the infimum over $s>0$ yields $-\infty$. The application of the Minimax Theorem is legitimate at this level because of concavity in ζ and convexity in $s>0$. A more streamlined derivation of this bound is to simply note that $-\log\Pr_{S\sim\nu}[f(S)\geq t]=\mathscr{D}_{\mathrm{KL}}\left(\frac{\nu\cdot\mathbf{1}_{f\geq t}}{\Pr_{S\sim\nu}[f(S)\geq t]}\,\|\,\nu\right)$. Restricting to product measures $\pi(\boldsymbol{m})$ and applying the correctness assumption for naïve mean-field yields the claim. □

Remark 1. The final claim regarding the mean-field assumption is not quite rigorous due to issues of swapping infimum and supremum. In particular, concavity is lost in $\mathcal{F}_\nu^{\mathsf{NMF}}(s\cdot f)$. Nonetheless, typically the proof technique establishing mean-field behavior also confirms Eq. (3). For instance, Theorem 2.1 below shows that existence of low-entropy measure decompositions implies that one does not lose much by restricting ζ to product measures (recall that low-entropy decompositions of a very similar flavor were used in a previous lecture to establish mean-field behavior).

In light of Lemma 1.1, the hope is that we can obtain sharp "function aware" control on the right-hand side of Eq. (3), perhaps up to additional additive losses of order $o(|\mathcal{U}|)$.

For intuition, note that an alternative interpretation of the approximation Eq. (3) is the following. Consider the conditioned measure $\nu_{f\geq t}(S)\propto\nu(S)\mathbf{1}[f(S)\geq t]$; as observed earlier, $\mathscr{D}_{\mathrm{KL}}(\nu_{f\geq t}\,\|\,\nu)=-\log\Pr_{S\sim\nu}[f(S)\geq t]$. Now even in the regime where Eq. (3) holds, $\nu_{f\geq t}$ need not be literally close to ν in KL-divergence, i.e. $\mathscr{D}_{\mathrm{KL}}(\nu_{f\geq t}\,\|\,\nu)\leq o(|\mathcal{U}|)$ can fail. This is simply because the Hamiltonian f itself can be used as a distinguisher for the pair $\nu_{f\geq t},\nu$. Indeed, since we're in the large deviations regime, if $\mathbb{E}_\nu[f]$ has order $|\mathcal{U}|$ and $t=(1+\delta)\mathbb{E}_\nu[f]$ for constant $\delta>0$, then $\left|\mathbb{E}_{\nu_{f\geq t}}[f]-\mathbb{E}_\nu[f]\right|\geq\delta\cdot\mathbb{E}_\nu[f]\geq\Omega(|\mathcal{U}|)$.

However, Eq. (3) says that w.r.t. the specific "test function" $\zeta\mapsto\mathscr{D}_{\mathrm{KL}}(\zeta\,\|\,\nu)$, $\nu_{f\geq t}$ is "indistinguishable" from *some* product measure $\pi(\boldsymbol{m})$ (i.e. some exponential tilt of ν) satisfying $\mathbb{E}_{\pi(\boldsymbol{m})}[f]\geq t$. In particular, what is true is that Eq. (3) is equivalent to $\nu_{f\geq t}$ being close in KL-divergence to *some* exponential tilt of ν (up to additive $\pm o(|\mathcal{U}|)$ error).

Let us now see an interesting application to homomorphism densities in random graphs.

1.1 Application: Triangle Counts in $G(n,p)$

Fix a vertex set $V=[n]$, let $\mathcal{U}=\binom{V}{2}$, and let $\nu=\nu_p$ be the p-biased measure $\nu_p(E)=p^{|E|}(1-p)^{|\binom{V}{2}\setminus E|}$ for each $E\subseteq\binom{V}{2}$. The random pair $G=(V,E)$ is the Erdös–Rényi random graph $G(n,p)$ with edge density p. Let

$$T(G)\overset{\mathsf{def}}{=}\#\left\{uvw\in\binom{V}{3}:uv,vw,uw\in E\right\}$$

be the triangle count in G. Clearly, by linearity of expectation, $\mathbb{E}_{E\sim\nu_p}[T(G)]=p^3\binom{n}{3}$. Letting $t=(1+\delta)\mathbb{E}_{E\sim\nu_p}[T(G)]$, we have the following large deviation bound, which was a longstanding open problem in the study of random graphs.

Theorem 1.2 ([Aug20]; building on [CD16; LZ17; Eld18])**.** *If $n^{-1/2}\log^4 n\ll p\ll 1$, then for every $\delta>0$,*

$$\Pr_{E\sim\nu_p}\left[T(G)\geq(1+\delta)p^3\binom{n}{3}\right]=\exp\left(-(1\pm o(1))\min\left\{\frac{\delta^{2/3}}{2},\frac{\delta}{3}\right\}n^2p^2\log\frac{1}{p}\right).$$

Remark 2. The problem in the full range $n^{-1}\log n\ll p\ll 1$ was recently resolved in [HMS22].

Remark 3. Note that this equality is really a simultaneous upper and lower bound, each of which is allowed $1 \pm o(1)$ multiplicative error in the exponent. In this lecture, we only focus on upper bounds for brevity.

Remark 4. If one were instead to apply the modified log-Sobolev inequality for Glauber dynamics for ν_p, which has constant $1/\binom{n}{2}$ just via standard entropy tensorization for product measures, then one would obtain something like

$$\Pr_{E \sim \nu_p}\left[T(G) \geq (1+\delta)p^3\binom{n}{3}\right] \leq \exp\left(-C\delta^2 n^2 p^6\right)$$

for some universal constant $C > 0$. This uses the fact that T is n-Lipschitz, since adding/deleting an edge uv can only affect the presence/absence of the triangles $\{uvw : w \in V\}$. Conceivably, one could optimize the constant $C > 0$, but the dependence on p and δ is not optimal.

We do not go into the full details of the proof here. The seminal papers [CD16; Eld18; EG18; Aus19; Aug20; Aug21] developed general-purpose techniques for making Lemma 1.1 applicable to problems like Theorem 1.2. For triangles in $G(n,p)$, and k-cliques more generally, the variational problem in Eq. (3) was then solved asymptotically in [LZ17; Bha+17]; note that they actually gave a solution to this optimization problem for all $1/n \ll p \ll 1$. We refer interested readers to [Bha+20] for additional applications to arithmetic progressions in random subsets of $\mathbb{Z}/n\mathbb{Z}$, and [CD16] for applications to exponential random graphs from social network analysis [LKR12].

2 On "Low-Complexity" Hamiltonians

In the rest of the lecture, we aim to give some indication of how to verify Eq. (3) holds. Our goal is to control

$$\mathcal{R}_\nu(f,t) \stackrel{\mathsf{def}}{=} \inf_{\zeta}\left\{\mathscr{D}_{\mathrm{KL}}\left(\zeta \,\|\, \nu\right) \,\middle|\, \mathbb{E}_\zeta[f] \geq t\right\}$$

by

$$\mathcal{R}_\nu^{\mathsf{NMF}}(f,t) \stackrel{\mathsf{def}}{=} \inf_{\boldsymbol{m} \in [0,1]^{\mathcal{U}}}\left\{\mathscr{D}_{\mathrm{KL}}\left(\pi(\boldsymbol{m}) \,\|\, \nu\right) \,\middle|\, \mathbb{E}_{\pi(\boldsymbol{m})}[f] \geq t\right\}$$

up to small additive error. The fundamental idea is again decomposition. Previous approaches, specialized to the random graph setting (see Section 1.1), leveraged various *regularity* results of Szemerédi-type to formalize this theme. In this lecture, we go via measure decompositions, following [Eld18; Aus19]. The following is a direct adaptation of a result we saw in the previous lecture on the naïve mean-field approximation, although with a more stringent condition for the component measures.

Theorem 2.1. *Let $f : 2^{\mathcal{U}} \to \mathbb{R}$ be some Hamiltonian, and ν be a product measure over $2^{\mathcal{U}}$. Suppose there are $\alpha, \eta > 0$ (possibly depending on $|\mathcal{U}|$) such that for every $s > 0$, we can decompose $\mu(S) \propto \nu(S)e^{s \cdot f(S)}$ as a mixture $\mathbb{E}_{\theta \sim \xi}\left[\mu^{(\theta)}\right]$, where ξ is a distribution over some auxiliary state space $\mathcal{I}$, and each component measure $\mu^{(\theta)}$ is again a distribution over $2^{\mathcal{U}}$, such that the following properties hold:*

- ***"Low Entropy" Mixture:*** $\mathbb{E}_{\theta \sim \xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(\theta)} \,\|\, \nu\right)\right] - \mathscr{D}_{\mathrm{KL}}\left(\mu \,\|\, \nu\right) \leq \alpha$.
- ***"Near-Product" Components:*** *With $1 - o(1)$ probability over $\theta \sim \xi$, we have $\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f] \leq \eta$ and $\mathbb{E}_{\mu^{(\theta)}}[f] \geq \mathbb{E}_\mu[f] - \tilde{O}(\eta)$. (Note that $\mathbb{E}_{\theta \sim \xi}\left[\mathbb{E}_{\mu^{(\theta)}}[f]\right] = \mathbb{E}_\mu[f]$.)*

Then

$$\mathcal{R}_\nu(f,t) \leq \mathcal{R}_\nu^{\mathsf{NMF}}(f,t) \leq \mathcal{R}_\nu(f, t + \tilde{O}(\eta)) + \alpha.$$

We give a proof in Appendix A. In the setting of large deviations, one should think of t as linear in dimension (e.g. $\binom{n}{2}$ in the $G(n,p)$ setting; see Section 1.1), while α, η are sublinear in dimension so that they can be viewed as $1 \pm o(1)$ multiplicative errors compared to t. The game then becomes to find such decompositions, which would allow one to invoke Lemma 1.1. The following is a pervasive theme in this line of research.

Theme 2.2. *If the Hamiltonian f satisfies a suitable "low-complexity" condition (e.g. as measured via its set of gradients), then such a decomposition in the style of Theorem 2.1 exists, and Lemma 1.1 is applicable.*

To illustrate this theme, we prove a decomposition theorem due to Austin [Aus19]. Throughout this section, we let $f : \{\pm 1\}^n \to \mathbb{R}$ be some fixed Hamiltonian over the Boolean cube, and let ν be some fixed reference *product* measure over $\{\pm 1\}^n$ (e.g. $\mathsf{Unif}\{\pm 1\}^n$). Recall that the *discrete gradient* for functions on $\{\pm 1\}^n$ is defined via

$$\nabla f(\sigma) = [\partial_i f(\sigma)]_{i=1}^n \in \mathbb{R}^n \qquad \text{where} \qquad \partial_i f(\sigma) = \frac{f(\sigma_{-i}, +1) - f(\sigma_{-i}, -1)}{2}. \tag{4}$$

Note that we could have instead defined $\partial_i f(\sigma) = \frac{f(\sigma) - f(\sigma^{\oplus i})}{2}$, but the above will be more convenient because $\partial_i f(\sigma)$ does not depend on σ_i. Along the lines of Theme 2.2, the resulting decomposition will be "good" if the collection of all gradients $\{\nabla f(\sigma) : \sigma \in \{\pm 1\}^n\}$ is "tame" in some sense.

Definition 1. *For a subset $S \subseteq \mathbb{R}^n$ and a parameter $\alpha > 0$, define $\mathsf{Cover}_\alpha(S)$ to be the smallest cardinality of any covering of S via subsets of $\mathbb{R}^n$ with ℓ_1-diameter at most α.*

Theorem 2.3 ([Aus19])**.** *Suppose f satisfies the following "low-complexity" guarantee for some parameters $\alpha, \eta > 0$ (possibly depending on n):*

$$\log \mathsf{Cover}_\alpha\left(\{\nabla f(\sigma) : \sigma \in \{\pm 1\}^n\}\right) \leq \eta. \tag{5}$$

Then the Gibbs measure $\mu(\sigma) \propto \nu(\sigma) e^{f(\sigma)}$ admits a decomposition $\mathbb{E}_{\theta \sim \xi}\left[\mu^{(\theta)}\right]$ such that $H(\xi) \leq \eta$, and for some collection of product measures $\{\pi^{(\theta)} : \theta \in \mathrm{supp}(\xi)\}$,

$$\mathbb{E}_{\theta \sim \xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(\theta)} \,\|\, \pi^{(\theta)}\right)\right] < \alpha + \eta.$$

Remark 5. The final inequality is very useful as it allows one to control deviations for all Lipschitz test functions simultaneously, including f itself. In particular, by Marton's transport-entropy inequality for product measures (see e.g. [GL10]; this is also a consequence of the modified log-Sobolev inequality), we have

$$\mathbb{E}_{\theta \sim \xi}\left[\mathscr{W}_1\left(\mu^{(\theta)}, \pi^{(\theta)}\right)\right] < \alpha + \eta,$$

where this Wasserstein distance is defined w.r.t. the Hamming metric on $\{\pm 1\}^n$.

2.1 Austin's Approach via Dual Total Correlation

Perhaps motivated by a deeper examination of the (modified) log-Sobolev approach to large deviations (see e.g. Remark 4), one approach towards Definition 1 is to exponentially tilt the reference product measure ν in the direction of the various discrete gradients. This is related to the (modified) log-Sobolev inequality because the discrete gradients appear directly in the Dirichlet form of Glauber dynamics, and as we will see in the proof, we will take advantage of the (modified) log-Sobolev inequality for product measures. To formalize this, for any other probability measure μ over $\{\pm 1\}^n$, define the *dual total correlation* [Han75] by

$$\mathsf{DTC}(\mu) \overset{\mathsf{def}}{=} \sum_{i=1}^n \mathbb{E}_{\tau \sim \mu_{-i}}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu_i^\tau \,\|\, \nu_i\right)\right] - \mathscr{D}_{\mathrm{KL}}\left(\mu \,\|\, \nu\right). \tag{6}$$

This quantity enjoys the following nice properties.

Fact 2.4 (Properties of DTC)**.**

- $\mathsf{DTC}(\mu) \geq 0$ *for every μ.*
- *The definition of $\mathsf{DTC}(\mu)$ is independent of the choice of reference product measure ν.*
- $\mathsf{DTC}(\mu)$ *may be alternatively expressed as*

$$\mathsf{DTC}(\mu) = \mathbb{E}_{\sigma \sim \mu}\left[\mathscr{D}_{\mathrm{KL}}\left(\pi^{(\sigma)} \,\|\, \nu\right)\right] - \mathscr{D}_{\mathrm{KL}}\left(\mu \,\|\, \nu\right),$$

where $\pi^{(\sigma)}$ is the unique product measure over $\{\pm 1\}^n$ with marginals $\pi_i^{(\sigma)} = \mu_i^{\sigma_{-i}}$ for all $\sigma \in \{\pm 1\}^n$. Note that $\pi^{(\sigma)} = \mathcal{T}_{\nabla f(\sigma)} \nu$.

The first claim is just approximate tensorization of entropy for product measures, which can be proved inductively via standard methods. It is more or less the log-Sobolev inequality for product measures. The second and third claims can be verified by direct calculation. One should think of the first term in $\mathsf{DTC}(\mu)$ as being the Dirichlet form of Glauber dynamics for ν evaluated at the (log-)density of μ w.r.t. ν. We omit the proof for brevity.

The following is the main technical result of [Aus19].

Theorem 2.5 (Main Technical; [Aus19])**.** *Suppose there is an $\alpha > 0$ (possibly depending on n) and a partition $\mathcal{P}$ of $\{\pm 1\}^n$ satisfying the following condition:*

$$\|\nabla f(\sigma) - \nabla f(\sigma')\|_1 < \alpha, \qquad \forall \sigma, \sigma' \text{ in the same part of } \mathcal{P}. \tag{7}$$

Let ξ be the mixture measure on $\mathcal{P}$ induced by μ, i.e. $\xi(P) = \mu(P)$ for each part $P \in \mathcal{P}$, and let the component measures be given by conditioning, i.e. $\mu^{(P)} = \mu \mid P$ for each part $P \in \mathcal{P}$. Then the following estimate holds:

$$\mathsf{DTC}(\mu) + \mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(P)}}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(P)} \,\|\, \pi^{(\sigma)}\right)\right]\right] < H(\xi) + \alpha. \tag{8}$$

Let us first use this to prove Theorem 2.3.

Proof of Theorem 2.3. We take our decomposition to be the one furnished by Theorem 2.5 for an appropriate choice of $\mathcal{P}$. By the assumed metric entropy bound Eq. (5), there is a partition $\mathcal{Q}$ of the collection of discrete gradients $\{\nabla f(\sigma) : \sigma \in \{\pm 1\}^n\}$ into at most e^η-many sets of ℓ_1-diameter at most α. Now let $\mathcal{P} = \left\{(\nabla f)^{-1}(Q) : Q \in \mathcal{Q}\right\}$ be the partition of $\{\pm 1\}^n$ induced by pulling back the partition $\mathcal{Q}$ w.r.t. the map $\sigma \mapsto \nabla f(\sigma)$. We then take the product measure $\pi^{(\theta)}$ for σ minimizing $\mathscr{D}_{\mathrm{KL}}\left(\mu^{(\theta)} \,\|\, \pi^{(\sigma)}\right)$. The bound $H(\xi) \le \eta$ just follows from the cardinality bound $|\mathcal{P}| \le e^\eta$ and the Maximum Entropy Principle, while the second follows from the conclusion of Theorem 2.5 and nonnegativitiy of $\mathsf{DTC}(\mu)$ (see Fact 2.4). □

Remark 6. As observed in [Aus19], interestingly, Eq. (8) also implies that $\mathsf{DTC}(\mu) < H(\xi) + \alpha$. Since $\mathsf{DTC}(\mu)$ is essentially the *deficit* in the log-Sobolev inequality for the input test function f w.r.t. ν, this bound on $\mathsf{DTC}(\mu)$ essentially says that "low-complexity Hamiltonians f" nearly saturate the log-Sobolev inequality. This is what Eldan refers to as a *reverse log-Sobolev inequality* [Eld18]; see also [EL20] for the Gaussian case. This is also intimately related to *stability estimates* for the log-Sobolev inequality (see e.g. [ELS20]).

2.2 Proof of Theorem 2.5

We will need the following modified chain rule for KL-divergence, which is a straightforward consequence of the usual one. We omit the proof for brevity.

Lemma 2.6 (Modified Chain Rule)**.** *In the setting of Theorem 2.5, we have the identity*

$$\mathscr{D}_{\mathrm{KL}}(\mu \,\|\, \nu) = -H(\xi) + \mathbb{E}_{P\sim\xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(P)} \,\|\, \nu\right)\right].$$

For a fixed part of the partition $P \in \mathcal{P}$ and a fixed $\sigma \in \{\pm 1\}^n$, we have

$$\begin{aligned}\mathscr{D}_{\mathrm{KL}}\left(\mu^{(P)} \,\|\, \pi^{(\sigma)}\right) = &\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(P)} \,\|\, \nu\right) - \mathscr{D}_{\mathrm{KL}}\left(\pi^{(\sigma)} \,\|\, \nu\right)\right] \\ &- \left[\mathbb{E}_{\sigma'\sim\mu^{(P)}}[\langle\nabla f(\sigma), \sigma'\rangle] - \mathbb{E}_{\sigma'\sim\pi^{(\sigma)}}[\langle\nabla f(\sigma), \sigma'\rangle]\right]\end{aligned}$$

by using the standard "change of measure trick" for KL-divergence. Averaging over $\sigma \sim \mu^{(P)}$, and then averaging again over $P \sim \xi$, we obtain the following identity (after several applications of the Law of Total Expectation)

$$\begin{aligned}&\mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(P)}}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(P)} \,\|\, \pi^{(\sigma)}\right)\right]\right] \\ &\quad = \underbrace{\mathbb{E}_{P\sim\xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(P)} \,\|\, \nu\right)\right] - \mathbb{E}_{\sigma\sim\mu}\left[\mathscr{D}_{\mathrm{KL}}\left(\pi^{(\sigma)} \,\|\, \nu\right)\right]}_{(A)} \\ &\qquad - \underbrace{\mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma,\sigma'\sim\mu^{(P)}}[\langle\nabla f(\sigma), \sigma'\rangle] - \mathbb{E}_{\sigma\sim\mu^{(P)},\sigma'\sim\pi^{(\sigma)}}[\langle\nabla f(\sigma), \sigma'\rangle]\right]}_{(B)}.\end{aligned}$$

The left-hand side here is precisely the quantity in the left-hand side of Eq. (8) (except without the $\mathsf{DTC}(\mu)$ term). We bound (A) and (B) separately. The entropy term (A) can be rewritten as $(A) = H(\xi) - \mathsf{DTC}(\mu)$, which follows from Fact 2.4 and Lemma 2.6 by adding and subtracting $\mathscr{D}_{\mathrm{KL}}(\mu \,\|\, \nu)$. (B) is already a difference of quantities involving gradients and so it is conceivable we can apply our assumption Eq. (7). To formalize this, we claim that the following somewhat curious identity holds for the second term in (B):

$$\mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(P)},\sigma'\sim\pi^{(\sigma)}}[\langle\nabla f(\sigma),\sigma'\rangle]\right] = \mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(P)}}\left[\langle\nabla f(\sigma),\sigma\rangle\right]\right]. \tag{9}$$

Assuming the veracity of this identity, we have

$$\begin{aligned}(B) &= \mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma,\sigma'\sim\mu^{(P)}}\left[\langle\nabla f(\sigma)-\nabla f(\sigma'),\sigma'\rangle\right]\right] \\ &\le \mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma,\sigma'\sim\mu^{(P)}}\left[\|\nabla f(\sigma)-\nabla f(\sigma')\|_1 \cdot \|\sigma'\|_\infty\right]\right] && \text{(Hölder's Inequality)} \\ &\le \alpha. && \text{(By Eq. (7), since } \sigma,\sigma' \in P \text{ for some part } P \in \mathcal{P}\text{)}\end{aligned}$$

Putting this together with the previous observation on (A), the theorem follows. All that remains is to justify Eq. (9). The key is to observe that for each individual coordinate $i \in [n]$, we have

$$\mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(P)},\sigma'\sim\pi^{(\sigma)}}[\partial_i f(\sigma)\cdot\sigma_i']\right] = \mathbb{E}_{P\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(P)}}\left[\partial_i f(\sigma)\cdot\sigma_i\right]\right]$$

just by using $\pi_i^{(\sigma)} = \mu_i^{\sigma_{-i}}$ and the fact that $\partial_i f(\sigma)$ is independent of σ_i. Indeed, in both sides, the overarching law of both σ and (σ_{-i},σ_i') is $\mu = \mathbb{E}_{P\sim\xi}\left[\mu^{(P)}\right]$, and we may replace $\partial_i f(\sigma)$ with $\partial_i f(\sigma')$ in the left-hand side, for instance. Summing over all $i \in [n]$ and using linearity of expectation yields Eq. (9), and so we are done.

References

[Aug20] Fanny Augeri. "Nonlinear large deviation bounds with applications to Wigner matrices and sparse Erdös–Rényi graphs". In: *The Annals of Probability* 48.5 (2020), pp. 2404–2448 (cit. on pp. 2, 3).

[Aug21] Fanny Augeri. "A transportation approach to the mean-field approximation". In: *Probability Theory and Related Fields* 180 (2021), pp. 1–32 (cit. on p. 3).

[Aus19] Tim Austin. "The structure of low-complexity Gibbs measures on product spaces". In: *The Annals of Probability* 47.6 (2019), pp. 4002–4023 (cit. on pp. 3–5).

[Bha+17] Bhaswar B. Bhattacharya, Shirshendu Ganguly, Eyal Lubetzky, and Yufei Zhao. "Upper tails and independence polynomials in random graphs". In: *Advances in Mathematics* 319 (2017), pp. 313–347. ISSN: 0001-8708. DOI: https://doi.org/10.1016/j.aim.2017.08.003 (cit. on p. 3).

[Bha+20] Bhaswar B Bhattacharya, Shirshendu Ganguly, Xuancheng Shao, and Yufei Zhao. "Upper Tail Large Deviations for Arithmetic Progressions in a Random Set". In: *International Mathematics Research Notices* 2020.1 (2020), pp. 167–213 (cit. on p. 3).

[CD16] Sourav Chatterjee and Amir Dembo. "Nonlinear large deviations". In: *Advances in Mathematics* 299 (2016), pp. 396–450. ISSN: 0001-8708. DOI: https://doi.org/10.1016/j.aim.2016.05.017 (cit. on pp. 2, 3).

[EG18] Ronen Eldan and Renan Gross. "Decomposition of mean-field Gibbs distributions into product measures". In: *Electronic Journal of Probability* 23 (2018), pp. 1–24 (cit. on p. 3).

[EL20] Ronen Eldan and Michel Ledoux. "A Dimension-Free Reverse Logarithmic Sobolev Inequality for Low-Complexity Functions in Gaussian Space". In: *Geometric Aspects of Functional Analysis. Lecture Notes in Mathematics, vol 2256* (2020), pp. 263–271 (cit. on p. 5).

[Eld18] Ronen Eldan. "Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations". In: *Geometric and Functional Analysis* 28 (2018), pp. 1548–1596 (cit. on pp. 2, 3, 5).

[ELS20] Ronen Eldan, Joseph Lehec, and Yair Shenfeld. "Stability of the logarithmic Sobolev inequality via the Föllmer process". In: *Annales de l'Institut Henri Poincaré – Probabilités et Statistiques* 56.3 (2020), pp. 2253–2269 (cit. on p. 5).

[GL10] Nathael Gozlan and Christian Léonard. “Transport Inequalities. A Survey”. In: *Markov Processes And Related Fields* 16 (4 2010), pp. 635–736 (cit. on p. 4).

[Han75] Te Sun Han. “Linear dependence structure of the entropy space”. In: *Information and Control* 29.4 (1975), pp. 337–368. ISSN: 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(75)80004-0 (cit. on p. 4).

[HMS22] Matan Harel, Frank Mousset, and Wojciech Samotij. “Upper tails via high moments and entropic stability”. In: *Duke Mathematical Journal* 171.10 (2022), pp. 2089–2192 (cit. on p. 2).

[LKR12] Dean Lusher, Johan Koskinen, and Garry Robins. *Exponential Random Graph Models for Social Networks Theory, Methods, and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, 2012 (cit. on p. 3).

[LZ17] Eyal Lubetzky and Yufei Zhao. “On the Variational Problem for Upper Tails in Sparse Random Graphs”. In: *Random Struct. Algorithms* 50.3 (May 2017), pp. 420–436. ISSN: 1042-9832. DOI: 10.1002/rsa.20658 (cit. on pp. 2, 3).

A Unfinished Proofs

Proof of Theorem 2.1. The proof is similar to before, except complicated by the hard constraint that $\mathbb{E}_\zeta[f]$ must exceed some threshold. The lower bound is trivial. For the upper bound, observe that by the Maximum Entropy Principle, the variational problem for $\mathcal{R}_\nu(f, t + \tilde{O}(\eta))$ is attained by the Gibbs measure $\mu(S) \propto \nu(S)e^{s \cdot f(S)}$ for some $s > 0$. Applying this decomposition, we have

$$\begin{aligned}
\mathcal{R}_\nu(f, t + \tilde{O}(\eta)) &= \mathscr{D}_{\mathrm{KL}}(\mu \,\|\, \nu) && \text{(Optimality of } \mu\text{)}\\
&= \underbrace{\mathscr{D}_{\mathrm{KL}}(\mu \,\|\, \nu) - \mathbb{E}_{\theta\sim\xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(\theta)} \,\|\, \nu\right)\right]}_{\geq -\alpha} + \mathbb{E}_{\theta\sim\xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\mu^{(\theta)} \,\|\, \nu\right)\right]\\
&\geq -\alpha + \mathbb{E}_{\theta\sim\xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\pi\left(\mu^{(\theta)}\right) \,\|\, \nu\right)\right]\\
& && \text{(Using } \mathscr{D}_{\mathrm{KL}}\left(\mu^{(\theta)} \,\|\, \nu\right) \geq \mathscr{D}_{\mathrm{KL}}\left(\pi\left(\mu^{(\theta)}\right) \,\|\, \nu\right) \text{ since } \nu \text{ is product)}\\
&\geq \Pr_{\theta\sim\xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\pi\left(\mu^{(\theta)}\right) \,\|\, \nu\right) \geq \mathcal{R}_\nu^{\mathsf{NMF}}(f,t)\right] \cdot \mathcal{R}_\nu^{\mathsf{NMF}}(f,t). && \text{(Markov's Inequality)}
\end{aligned}$$

Now we use our second assumption on the component measures to lower bound the probability that $\mathscr{D}_{\mathrm{KL}}\left(\pi\left(\mu^{(\theta)}\right) \,\|\, \nu\right)$ exceeds $\mathcal{R}_\nu^{\mathsf{NMF}}(f,t)$. Observe that

$$\begin{aligned}
&\Pr_{\theta\sim\xi}\left[\mathscr{D}_{\mathrm{KL}}\left(\pi\left(\mu^{(\theta)}\right) \,\|\, \nu\right) \geq \mathcal{R}_\nu^{\mathsf{NMF}}(f,t)\right]\\
&\geq \Pr_{\theta\sim\xi}\left[\mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f] \geq t\right]\\
&\geq \Pr_{\theta\sim\xi}\left[\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f] \leq \eta \,\text{and}\, \mathbb{E}_{\mu^{(\theta)}}[f] \geq t + \eta\right]\\
&\geq 1 - o(1). && \text{(Using the “near-product” assumption and } \mathbb{E}_\mu[f] \geq t + \tilde{O}(\eta)\text{)}
\end{aligned}$$

□