6.S891 Lecture 15: Entropic Independence (Continued), Spectral Independence from Mixing

Kuikui Liu

November 2, 2023

In this lecture, we wrap up our discussion of entropic independence by bringing in the lens of convex analysis. In the second half, we show that fast mixing of local Markov chains like Glauber dynamics implies spectral independence, giving a weak kind of converse to the local-to-global theorems we've already seen.

1 Entropic Independence via Spectral Independence for Tilts

We again restrict attention to distributions μ on $\{\pm 1\}^n$ for simplicity. Recall that η-entropic independence for μ means

$$\mathscr{D}_{\mathrm{KL}}(\nu_1 \,\|\, \mu_1) \le \frac{1+\eta}{n} \cdot \mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu), \qquad \forall \text{ distributions } \nu \text{ on } \{\pm 1\}^n. \tag{1}$$

Here, μ_1, ν_1 denote the induced marginal distributions on $[n] \times \{\pm 1\}$: $\mu_1(i,s) = \frac{1}{n} \Pr_{\sigma \sim \mu}[\sigma(i) = s]$. In the previous lecture, we saw how to deduce entropic independence from spectral independence for all pinnings plus marginal boundedness. In this lecture, we remove the marginal boundedness assumption, but at the cost of a significantly stronger our spectral independence requirement. This alternative set of hypotheses is crucial for certain applications (e.g. determinantal point processes), where marginal boundedness fails dramatically.

Definition 1 (Exponential Tilt)**.** *For a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ and a distribution μ on $\{\pm 1\}^n$, define the* exponential tilt $\mathcal{T}_{\boldsymbol{\theta}}\mu$ *as the distribution on $\{\pm 1\}^n$ given by*

$$(\mathcal{T}_{\boldsymbol{\theta}}\mu)(\sigma) \propto \mu(\sigma) \cdot \exp\left(\langle \boldsymbol{\theta}, \sigma \rangle\right), \qquad \forall \sigma \in \{\pm 1\}^n. \tag{2}$$

Exponential tilts are natural distributions from the perspective of the *maximum entropy principle*, which we explain in a moment. In the language of statistical physics, the vector v induces an *external field* on the coordinates of $\sigma \sim \mu$. We have the following theorem.

Theorem 1.1 ([Ana+22; CE22])**.** *Let μ be a probability measure on $\{\pm 1\}^n$, and fix a parameter η. Then the following are equivalent:*

- *For every $\boldsymbol{\theta} \in \mathbb{R}^n$, the tilted measure $\mathcal{T}_{\boldsymbol{\theta}}\mu$ is η-spectrally independent.*
- *For every $\boldsymbol{\theta} \in \mathbb{R}^n$, the tilted measure $\mathcal{T}_{\boldsymbol{\theta}}\mu$ is η-entropically independent.*

By sending the entries of $\boldsymbol{\theta}$ to $\pm\infty$, one can obtain all conditional measures of μ as special cases of exponential tilts. Hence, the spectral independence assumption in Theorem 1.1 is genuinely stronger than spectral independence for all pinnings. However, we have made no marginal boundedness assumptions.

1.1 Exponential Tilts as Maximum Entropy Distributions

Let us now elucidate why exponential tilts appear in Theorem 1.1.

Fact 1.2 (Special Case of Theorem 1.4)**.** *Let $\boldsymbol{q}$ be a probability measure on $[n] \times \{\pm 1\}$. Then the variational problem*

$$\begin{aligned} &\inf_{\nu} \mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu) \\ s.t. \quad &\nu_1 = \boldsymbol{q} \end{aligned}$$

over distributions ν on $\{\pm 1\}^n$ is optimized by an exponential tilt $\mathcal{T}_{\boldsymbol{\theta}}\mu$ for some $\boldsymbol{\theta} \in \mathbb{R}^n$.

We prove Fact 1.2 in greater generality later. But the key point here is that to certify η-entropic independence, it suffices to restrict attention to exponential tilts $\nu = \mathcal{T}_{\boldsymbol{\theta}}\mu$ in Eq. (1). This is convenient because we can simultaneously encapsulate all exponential tilts in a single function:

$$\mathcal{L}_\mu(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \mathbb{E}_{\sigma\sim\mu}\left[\exp\left(\langle \boldsymbol{\theta}, \sigma\rangle\right)\right], \qquad \forall \boldsymbol{\theta} \in \mathbb{R}^n. \tag{3}$$

This is sometimes called the *cumulant generating function* of μ, or the *logarithmic Laplace transform* of μ. It is the logarithm of the moment generating function of μ. As the name suggests, its derivatives capture moment information of any exponential tilt of μ. Note that this is essentially the same as the log-partition function $\log Z(\boldsymbol{\lambda})$ we saw in various contexts in prior lectures, where $Z(\boldsymbol{\lambda})$ is viewed as a multivariate polynomial; one just does a change of variables e.g. $\boldsymbol{\lambda} \propto e^{2\boldsymbol{\theta}}$.

Theorem 1.3 (Special Case of Theorem 1.6)**.** *Let μ be a probability measure on $\{\pm 1\}^n$. Then $\mathcal{L}_\mu(\cdot)$ has the following properties:*

1. *It is smooth and strictly convex.*
2. *Its gradient gives the mean of $\mathcal{T}_{\boldsymbol{\theta}}\mu$: $\nabla \mathcal{L}_\mu(\boldsymbol{\theta}) = \mathbb{E}_{\sigma\sim\mathcal{T}_{\boldsymbol{\theta}}\mu}[\sigma]$.*
3. *Its Hessian gives the covariance of $\mathcal{T}_{\boldsymbol{\theta}}\mu$: $\nabla^2 \mathcal{L}_\mu(\boldsymbol{\theta}) = \mathrm{Cov}\left(\mathcal{T}_{\boldsymbol{\theta}}\mu\right)$.*
4. *Its* convex conjugate *(or* Legendre transform/Fenchel dual*) $\mathcal{L}^*_\mu(\boldsymbol{m}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}}\left\{\langle \boldsymbol{\theta}, \boldsymbol{m}\rangle - \mathcal{L}_\mu(\boldsymbol{\theta})\right\}$ has the formula*

$$\mathcal{L}^*_\mu(\boldsymbol{m}) = \mathscr{D}_{\mathrm{KL}}\left(\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m})}\mu \,\|\, \mu\right), \tag{4}$$

where $\boldsymbol{\theta}^*(\boldsymbol{m}) = \nabla \mathcal{L}^*_\mu(\boldsymbol{m})$ is the optimizer in the definition of $\mathcal{L}^*_\mu(\boldsymbol{m})$. Furthermore, $\nabla \mathcal{L}_\mu(\cdot)$ and $\nabla \mathcal{L}^*_\mu(\cdot)$ are inverses of each other as maps from $\mathbb{R}^n$ to $\mathbb{R}^n$, and $\nabla^2 \mathcal{L}^*_\mu(\boldsymbol{m}) = \mathrm{Cov}\left(\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m})}\mu\right)^{-1}$.

We also prove this in greater generality later. For now, we use it to connect entropic independence with spectral independence.

Proof of Theorem 1.1. We first show spectral independence for all tilts implies entropic independence for all tilts. Without loss of generality, we verify entropic independence for μ itself. Let us first translate μ_1 into the language of mean vectors $\boldsymbol{m}(\mu) \stackrel{\text{def}}{=} \mathbb{E}_{\sigma\sim\mu}[\sigma]$ so that we can use Theorem 1.3. Observe that $\mu_1(i,+1) = \frac{1}{n}\cdot\frac{1+m_i(\mu)}{2}, \mu_1(i,-1) = \frac{1}{n}\cdot\frac{1-m_i(\mu)}{2}$. It follows that

$$\begin{aligned}
\mathscr{D}_{\mathrm{KL}}\left(\nu_1 \,\|\, \mu_1\right) &= \sum_{i=1}^n \left(\nu_1(i,+1)\log\frac{\nu_1(i,+1)}{\mu_1(i,+1)} + \nu_1(i,-1)\log\frac{\nu_1(i,-1)}{\mu_1(i,-1)}\right) && \text{(Definition)}\\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{1+m_i(\nu)}{2}\log\frac{1+m_i(\nu)}{1+m_i(\mu)} + \frac{1-m_i(\nu)}{2}\log\frac{1-m_i(\nu)}{1-m_i(\mu)}\right)\\
&\stackrel{\text{def}}{=} \frac{1}{n}\Phi_\mu\left(\boldsymbol{m}(\nu)\right),
\end{aligned}$$

where $\Phi_\mu(\boldsymbol{m})$ is a function on $[-1,1]^n$ which isolates the dependence on the mean of ν. Hence, entropic independence is equivalent to showing

$$\Phi_\mu\left(\boldsymbol{m}(\nu)\right) \le (1+\eta)\cdot\mathscr{D}_{\mathrm{KL}}(\nu\,\|\,\mu), \qquad \forall \nu.$$

By Fact 1.2, it suffices to verify this inequality for measures of the form $\nu = \mathcal{T}_{\boldsymbol{\theta}}\mu$ for any $\boldsymbol{\theta} \in \mathbb{R}^n$, which by Theorem 1.3, is equivalent to

$$\mathcal{F}_\mu(\boldsymbol{m}) \stackrel{\text{def}}{=} (1+\eta)\cdot\mathcal{L}^*_\mu(\boldsymbol{m}) - \Phi_\mu(\boldsymbol{m}) \ge 0, \qquad \forall \boldsymbol{m} \in [-1,1]^n. \tag{5}$$

To do this, we show that $\mathcal{F}_\mu(\boldsymbol{m})$ is convex, and that at some point $\boldsymbol{m}^*$, $\mathcal{F}_\mu(\boldsymbol{m}^*) \ge 0$ and $\nabla\mathcal{F}_\mu(\boldsymbol{m}^*) = 0$. This is indeed sufficient, since convexity implies $\mathcal{F}_\mu(\boldsymbol{m})$ is lower bounded by its first-order Taylor approximation around $\boldsymbol{m}^*$, which is some nonnegative constant.

Let $\boldsymbol{m}^* = \boldsymbol{m}(\mu) = \nabla\mathcal{L}_\mu(\boldsymbol{0})$. Then $\mathcal{F}_\mu(\boldsymbol{m^*}) = (1+\eta)\cdot\mathscr{D}_{\mathrm{KL}}(\mu\,\|\,\mu) - \Phi_\mu(\boldsymbol{m}(\mu)) = 0$. Furthermore, $\nabla\mathcal{L}^*_\mu(\boldsymbol{m}^*) = \boldsymbol{0}$ since $\nabla\mathcal{L}^*_\mu(\cdot)$ and $\nabla\mathcal{L}_\mu(\cdot)$ are inverse maps, and $\nabla\Phi_\mu(\boldsymbol{m}^*) = \boldsymbol{0}$ since

$$\nabla\Phi_\mu(\boldsymbol{m}) = \frac{1}{2}\log\frac{1-\boldsymbol{m}(\mu)}{1+\boldsymbol{m}(\mu)} - \frac{1}{2}\log\frac{1-\boldsymbol{m}}{1+\boldsymbol{m}},$$

where the function is applied entrywise. Let us now show the Hessian is positive semidefinite. On the one hand, $\nabla^2 \mathcal{L}^*_\mu(\boldsymbol{m}) = \mathrm{Cov}\left(\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m})}\mu\right)^{-1}$ by Theorem 1.3. On the other hand, $\nabla^2 \Phi_\mu(\boldsymbol{m}) = \mathrm{diag}\left(1-\boldsymbol{m}^2\right)^{-1}$. Hence, convexity of $\mathcal{F}_\mu(\boldsymbol{m})$ is equivalent to

$$\mathrm{Cov}\left(\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m})}\mu\right) \preceq (1+\eta)\cdot \mathrm{diag}(1-\boldsymbol{m}^2) = (1+\eta)\cdot \mathrm{diag}\left(\mathrm{Var}_{\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m})}\mu}(\sigma_i)\right)_{i\in[n]}.$$

In the final step, we used that the mean of $\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m})}\mu$ is $\boldsymbol{m}$ by Theorem 1.3. This matrix inequality is exactly η-spectral independence of $\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m})}\mu$. By assumption, this holds for arbitrary $\boldsymbol{m}\in[-1,1]^n$, and so we're done.

Now assume entropic independence holds for all tilts; we wish to deduce spectral independence for all tilts. In other words, by the above calculations, our assumption is that $\mathcal{F}_{\mathcal{T}_{\theta}\mu}(\boldsymbol{m}) \geq 0$ for every $\boldsymbol{m}\in[-1,1]^n, \boldsymbol{\theta}\in\mathbb{R}^n$, and our desired conclusion is global convexity of $\mathcal{F}_\mu(\cdot)$, i.e. $\nabla^2\mathcal{F}_\mu(\boldsymbol{m}) \succeq 0$ for all $\boldsymbol{m}\in[-1,1]^n$. The key is to observe that

$$\mathcal{F}_\mu(\boldsymbol{m}) - \mathcal{F}_\mu(\boldsymbol{m}^*) - \langle \nabla\mathcal{F}_\mu(\boldsymbol{m}^*), \boldsymbol{m}-\boldsymbol{m}^*\rangle = \mathcal{F}_{\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m}^*)}\mu}(\boldsymbol{m}) \geq 0,$$

where the first equality follows by direct calculation (see Appendix B), and the second inequality follows by assumption. This is exactly the first-order characterization of global convexity applied to $\mathcal{F}_\mu(\cdot)$, and so we're done. □

1.2 The Maximum Entropy Principle

Fact 1.2 is a special case of a much more general result on maximum entropy distributions. We sketch of the proof is provided in Appendix A.

Theorem 1.4 (Maximum Entropy Distribution; see e.g. [WJ08])**.** *Let μ be some base probability measure on a (finite) state space Ω. Let $\boldsymbol{\varphi}:\Omega\to\mathbb{R}^d$ be a collection of d real-valued functions and $\boldsymbol{m}\in\mathbb{R}^d$. Then the solution to the following variational problem*[1]

$$\begin{aligned} &\inf_\nu \mathscr{D}_{\mathrm{KL}}(\nu\,\|\,\mu) \\ s.t.\quad &\mathbb{E}_{x\sim\nu}\left[\boldsymbol{\varphi}(x)\right] = \boldsymbol{m}, \end{aligned} \tag{6}$$

over probability measures ν has the form

$$\mu_{\boldsymbol{\theta}}(x) \propto \mu(x)\cdot\exp\left(\langle\boldsymbol{\theta},\boldsymbol{\varphi}(x)\rangle\right), \tag{7}$$

for some vector $\boldsymbol{\theta}\in\mathbb{R}^d$.

In statistics lingo, the collection $\boldsymbol{\varphi}$ are often called *sufficient statistics*, and the induced collection of distributions of the form Eq. (7) is called the *exponential family* associated to $\boldsymbol{\varphi}$. Every distribution we have come across is a maximum entropy distribution w.r.t. some very natural set of sufficient statistics $\boldsymbol{\varphi}$. For instance, Ising models arise by letting the base measure μ be uniform over $\{\pm1\}^n$, and letting $\boldsymbol{\varphi}$ be a collection of degree-2 polynomials (e.g. $\beta x_i x_j$ for edges $ij\in E$ in some underlying graph). For more in depth discussion, see the monograph [WJ08].

1.3 General Convex Analysis of KL-Divergence

Theorem 1.3 is also a special case of much more general convex duality phenomena. The following theorem will also be useful in future lectures on variational inference.

Theorem 1.5. *Let μ be a base probability measure on a (finite) state space Ω. Then the functions $f\mapsto \log\mathbb{E}_{x\sim\mu}\left[e^{f(x)}\right]$ and $\nu\mapsto\mathscr{D}_{\mathrm{KL}}(\nu\,\|\,\mu)$ are smooth and strictly convex.*[2] *Furthermore, we have the following duality relations between them.*

- ***Gibbs Variational Principle:*** *For every function $f:\Omega\to\mathbb{R}$,*

$$\textit{(Primal Program)}\qquad \log\mathbb{E}_{x\sim\mu}\left[e^{f(x)}\right] = \sup_\nu\left\{\mathbb{E}_{x\sim\nu}\left[f(x)\right] - \mathscr{D}_{\mathrm{KL}}(\nu\,\|\,\mu)\right\}. \tag{8}$$

Furthermore, the supremum is uniquely attained at the measure $\nu(x)\propto\mu(x)e^{f(x)}$.

[1] On the surface, the optimization in Eq. (6) doesn't "look like" a *maximum* entropy problem. But it is; indeed when μ is uniform over Ω, minimizing $\mathscr{D}_{\mathrm{KL}}(\nu\,\|\,\mu)$ is equivalent to maximizing Shannon entropy $H(\nu)$.

[2] Technically, $f\mapsto\log\mathbb{E}_{x\sim\mu}\left[e^{f(x)}\right]$ is not strictly convex if you allow shifting f by an additive constant, but this point is immaterial.

- ***Donsker–Varadhan Variational Representation:*** *For every probability measure ν on Ω,*

$$\textit{(Dual Program)} \qquad \mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu) = \sup_f \left\{ \mathbb{E}_{x\sim\nu}[f(x)] - \log \mathbb{E}_{x\sim\mu}\left[e^{f(x)}\right] \right\}. \tag{9}$$

Furthermore, the supremum is uniquely attained at the function $f(x) = \log \frac{\nu(x)}{\mu(x)}$ (up to shifting by an additive constant).

A proof is provided in Appendix A for completeness. For now, we combine it with Theorem 1.4 to deduce the following.

Theorem 1.6 (Cumulants and Entropy; see e.g. [WJ08])**.** *Let μ be some base probability measure on a (finite) state space Ω, and let $\boldsymbol{\varphi} : \Omega \to \mathbb{R}^m$ be a collection of d real-valued functions. Let*

$$\mathcal{L}_{\mu,\boldsymbol{\varphi}}(\boldsymbol{\theta}) \stackrel{\mathsf{def}}{=} \log \mathbb{E}_{x\sim\mu}\left[\exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x)\rangle\right)\right], \qquad \forall \boldsymbol{\theta} \in \mathbb{R}^d \tag{10}$$

denote the joint cumulant generating function *of $\boldsymbol{\varphi}$ w.r.t. μ. Then $\mathcal{L}_{\mu,\boldsymbol{\varphi}}(\cdot)$ has the following properties:*

1. *It is smooth and strictly convex.*
2. *Its gradient gives the mean of $\boldsymbol{\varphi}$ under $\mu_{\boldsymbol{\theta}}$ from Eq. (7): $\nabla \mathcal{L}_{\mu,\boldsymbol{\varphi}}(\boldsymbol{\theta}) = \mathbb{E}_{x\sim\mu_{\boldsymbol{\theta}}}[\boldsymbol{\varphi}(x)]$.*
3. *Its Hessian gives the covariance of $\boldsymbol{\varphi}$ under $\mu_{\boldsymbol{\theta}}$: $\nabla^2 \mathcal{L}_{\mu,\boldsymbol{\varphi}}(\boldsymbol{\theta}) = \mathbb{E}_{\mu_{\boldsymbol{\theta}}}\left[\boldsymbol{\varphi}^{\otimes 2}\right] - \mathbb{E}_{\mu_{\boldsymbol{\theta}}}[\boldsymbol{\varphi}]^{\otimes 2}$.*
4. *Its* convex conjugate *(or* Legendre transform/Fenchel dual*) $\mathcal{L}^*_{\mu,\boldsymbol{\varphi}}(\boldsymbol{m}) \stackrel{\mathsf{def}}{=} \sup_{\boldsymbol{\theta}} \{\langle \boldsymbol{\theta}, \boldsymbol{m}\rangle - \mathcal{L}_{\mu,\boldsymbol{\varphi}}(\boldsymbol{\theta})\}$ has the formula*

$$\mathcal{L}^*_{\mu,\boldsymbol{\varphi}}(\boldsymbol{m}) = \mathscr{D}_{\mathrm{KL}}\left(\mu_{\boldsymbol{\theta}^*(\boldsymbol{m})} \,\|\, \mu\right), \tag{11}$$

where $\boldsymbol{\theta}^*(\boldsymbol{m}) = \nabla \mathcal{L}^*_{\mu,\boldsymbol{\varphi}}(\boldsymbol{m})$ is the optimizer in the definition of $\mathcal{L}^*_{\mu,\boldsymbol{\varphi}}(\boldsymbol{m})$. Furthermore, $\nabla \mathcal{L}_{\mu,\boldsymbol{\varphi}}(\cdot)$ and $\nabla \mathcal{L}^*_{\mu,\boldsymbol{\varphi}}(\cdot)$ are inverses of each other as maps from $\mathbb{R}^d$ to $\mathbb{R}^d$, and $\nabla^2 \mathcal{L}^*_{\mu,\boldsymbol{\varphi}}(\boldsymbol{m}) = \mathrm{Cov}\left(\mu_{\boldsymbol{\theta}^*(\boldsymbol{m})}\right)^{-1}$.

Proof Sketch. The first three items are routine computations. For the final one, observe that

$$\begin{aligned}
\mathcal{L}_{\mu,\boldsymbol{\varphi}}(\boldsymbol{\theta}) &= \sup_{\nu} \{\langle \boldsymbol{\theta}, \mathbb{E}_{x\sim\nu}[\boldsymbol{\varphi}(x)]\rangle - \mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu)\} && \text{(Theorem 1.5)} \\
&= \sup_{\boldsymbol{\theta}'} \{\langle \boldsymbol{\theta}, \boldsymbol{m}(\mu_{\boldsymbol{\theta}'})\rangle - \mathscr{D}_{\mathrm{KL}}(\mu_{\boldsymbol{\theta}'} \,\|\, \mu)\} && \text{(Theorem 1.4)} \\
&= \sup_{\boldsymbol{m}} \left\{\langle \boldsymbol{\theta}, \boldsymbol{m}\rangle - \mathscr{D}_{\mathrm{KL}}\left(\mu_{\boldsymbol{\theta}(\boldsymbol{m})} \,\|\, \mu\right)\right\}, && \text{(Reparametrization over feasible } \boldsymbol{m})
\end{aligned}$$

where $\boldsymbol{\theta}(\boldsymbol{m})$ is the unique[3] $\boldsymbol{\theta}$ such that $\mathbb{E}_{x\sim\mu_{\boldsymbol{\theta}}}[\boldsymbol{\varphi}(x)] = \boldsymbol{m}$. It follows that the map $\boldsymbol{m} \mapsto \mathscr{D}_{\mathrm{KL}}\left(\mu_{\boldsymbol{\theta}(\boldsymbol{m})} \,\|\, \mu\right)$ must be the convex conjugate of $\mathcal{L}_{\mu,\boldsymbol{\varphi}}(\boldsymbol{\theta})$. The remaining claims plus the fact that $\boldsymbol{\theta}(\boldsymbol{m}) = \boldsymbol{\theta}^*(\boldsymbol{m})$ follow from standard arguments pertaining to convex conjugates. □

2 Spectral Independence from Optimal Spectral Gap

In the previous lectures, we proved that $O(1)$-spectral independence implies an inverse polynomial spectral gap for Glauber dynamics. Furthermore, in the setting of sparse graphical models, we can actually get the optimal $\Omega(1/n)$ spectral gap. Here, we prove a weak kind of converse. It says that $O(1)$-spectral independence is *necessary* for Glauber dynamics (or any other local Markov chain on $\{\pm 1\}^n$) to have the optimal $\Omega(1/n)$ spectral gap.

Lemma 2.1 ([Ana+23])**.** *Let μ be a probability measure over $\{\pm 1\}^n$, and suppose Glauber dynamics for μ satisfies the (optimal up to constants) Poincaré Inequality*

$$\mathrm{Var}_\mu(f) \le (1 + C)n \cdot \mathcal{E}_{\mathsf{GD}}(f, f), \qquad \forall f : \{\pm 1\}^n \to \mathbb{R},$$

for some constant $C \ge 0$. Then μ is C-spectrally independent.

[3] This requires justification, but for brevity, we sweep this under the rug.

Proof. We prove that $\mathrm{Cov}(\mu) \preceq (1+C)\cdot \mathrm{diag}(\mathrm{Var}_{\sigma\sim\mu}(\sigma_i))$. To do this, we need to show that for every vector $\boldsymbol{a}\in\mathbb{R}^n$, $\boldsymbol{a}^\top \mathrm{Cov}(\mu)\boldsymbol{a} \le (1+C)\cdot\sum_{i=1}^n \mathrm{Var}_\mu(\sigma_i)\cdot a_i^2$. Fix any such $\boldsymbol{a}\in\mathbb{R}^n$. We construct an appropriate test function $f_{\boldsymbol{a}}:\{\pm1\}^n\to\mathbb{R}$ to plug into the Poincaré Inequality.

Consider the *linear* test function $f_{\boldsymbol{a}}(\sigma) \stackrel{\mathsf{def}}{=} \langle \boldsymbol{a},\sigma\rangle$. Then

$$\begin{aligned}\mathrm{Var}_\mu\left(f_{\boldsymbol{a}}\right) &= \mathbb{E}_{\sigma\sim\mu}\left[\langle\boldsymbol{a},\sigma\rangle^2\right] - \mathbb{E}_{\sigma\sim\mu}\left[\langle\boldsymbol{a},\sigma\rangle\right]^2\\ &= \mathbb{E}_{\sigma\sim\mu}\left[\boldsymbol{a}^\top\sigma\sigma^\top\boldsymbol{a}\right] - \boldsymbol{a}^\top\mathbb{E}_{\sigma\sim\mu}[\sigma]\mathbb{E}_{\sigma\sim\mu}[\sigma]^\top\boldsymbol{a} = \boldsymbol{a}^\top\mathrm{Cov}(\mu)\boldsymbol{a}.\end{aligned}$$

On the other hand,

$$\begin{aligned}n\cdot\mathcal{E}_{\mathsf{GD}}(f_{\boldsymbol{a}},f_{\boldsymbol{a}}) &= \frac{n}{2}\sum_{\sigma\in\{\pm1\}^n}\mu(\sigma)\cdot\sum_{i=1}^n \mathsf{P}_{\mathsf{GD}}\left(\sigma\to\sigma^{\oplus i}\right)\cdot\left(f_{\boldsymbol{a}}(\sigma)-f_{\boldsymbol{a}}\left(x^{\oplus i}\right)\right)^2\\ &= 2\sum_{i=1}^n a_i^2\sum_{\sigma\in\{\pm1\}^n}\frac{\mu(\sigma)\mu\left(\sigma^{\oplus i}\right)}{\mu(\sigma)+\mu\left(\sigma^{\oplus i}\right)} \qquad \text{(Using } f_{\boldsymbol{a}}(\sigma)-f_{\boldsymbol{a}}\left(\sigma^{\oplus i}\right)=2a_i\sigma_i\text{)}\\ &= 4\sum_{i=1}^n a_i^2\sum_{\sigma:\sigma_i=+1}\left(\mu(\sigma)+\mu\left(\sigma^{\oplus i}\right)\right)\cdot h\left(\frac{\mu(\sigma)}{\mu(\sigma)+\mu\left(\sigma^{\oplus i}\right)}\right)\\ &\qquad\qquad \text{(Where } h(x)\stackrel{\mathsf{def}}{=}x(1-x)\text{)}\\ &\le 4\sum_{i=1}^n a_i^2\cdot h\left(\sum_{\sigma:\sigma_i=+1}\left(\mu(\sigma)+\mu\left(\sigma^{\oplus i}\right)\right)\cdot\frac{\mu(\sigma)}{\mu(\sigma)+\mu\left(\sigma^{\oplus i}\right)}\right)\\ &\qquad\qquad \text{(Jensen's Inequality and concavity of } h\text{)}\\ &= 4\sum_{i=1}^n a_i^2\cdot h\left(\mu_i(+1)\right)\\ &= \sum_{i=1}^n a_i^2\cdot\mathrm{Var}_\mu(\sigma_i).\end{aligned}$$

Put together, we have

$$\boldsymbol{a}^\top\mathrm{Cov}(\mu)\boldsymbol{a} = \mathrm{Var}_\mu\left(f_{\boldsymbol{a}}\right) \le (1+C)n\cdot\mathcal{E}_{\mathsf{GD}}(f_{\boldsymbol{a}},f_{\boldsymbol{a}}) \le (1+C)\cdot\boldsymbol{a}^\top\mathrm{diag}\left(\mathrm{Var}_{\sigma\sim\mu}(\sigma_i)\right)\boldsymbol{a}$$

as desired. □

3 ℓ_∞-Independence via Contractive Coupling

Our goal in this section is to give an analog of Lemma 2.1 where we strengthen both the assumption and the conclusion using couplings. We say μ is *ℓ_∞-independent with constant η* if $\|\Psi_\mu\|_{\ell_\infty\to\ell_\infty}\le 1+\eta$. We say μ is *η-coupling independent* if $\max_{i\in[n]}\mathscr{W}_1\left(\mu^{i\leftarrow+1},\mu^{i\leftarrow-1}\right)\le 1+\eta$, where $\mu^{i\leftarrow+1},\mu^{i\leftarrow-1}$ are viewed as distributions on $\{\pm1\}^n$. These terminology were first coined in [KKS21; CZ23], respectively. Here, $\mathscr{W}_1(\cdot,\cdot)$ denotes the 1-Wasserstein metric induced by Hamming distance $d_H(\cdot,\cdot)$ on $\{\pm1\}^n$.

We always have $\boldsymbol{\lambda}_{\max}(\Psi_\mu)\le\|\Psi_\mu\|_{\ell_\infty\to\ell_\infty}$ since the latter is a matrix norm induced by a vector norm. Hence, ℓ_∞-independence is stronger than spectral independence. [KKS21] established interesting Chernoff-type concentration inequalities which require ℓ_∞-independence. The following lemma shows that coupling independence is the strongest of the three.

Lemma 3.1 ([CZ23])**.** *For every $i\in[n]$,*

$$\sum_{j=1}^n|\Psi_\mu(i\to j)|\le\mathscr{W}_1\left(\mu^{i\leftarrow+1},\mu^{i\leftarrow-1}\right).$$

Proof. Observe that

$$\begin{aligned}
|\Psi_\mu(i \to j)| &= d_{\mathsf{TV}}\left(\mu_j^{i\leftarrow+1}, \mu_j^{i\leftarrow-1}\right) \\
&= \inf_{\substack{\text{Coupling } \xi_j \\ \text{of } \mu_j^{i\leftarrow+1}, \mu_j^{i\leftarrow-1}}} \mathbb{E}_{(x_j,y_j)\sim\xi_j}\left[\mathbb{I}\left[x_j \neq y_j\right]\right] && \text{(Coupling Lemma)} \\
&\leq \inf_{\substack{\text{Coupling } \xi \\ \text{of } \mu^{i\leftarrow+1}, \mu^{i\leftarrow-1}}} \mathbb{E}_{(x,y)\sim\xi}\left[\mathbb{I}\left[x_j \neq y_j\right]\right].
\end{aligned}$$

It follows that

$$\begin{aligned}
\sum_{j=1}^n |\Psi_\mu(i \to j)| &\leq \sum_{j=1}^n \inf_{\substack{\text{Coupling } \xi \\ \text{of } \mu^{i\leftarrow+1}, \mu^{i\leftarrow-1}}} \mathbb{E}_{(x,y)\sim\xi}\left[\mathbb{I}\left[x_j \neq y_j\right]\right] \\
&\leq \inf_{\substack{\text{Coupling } \xi \\ \text{of } \mu^{i\leftarrow+1}, \mu^{i\leftarrow-1}}} \underbrace{\sum_{j=1}^n \mathbb{E}_{(x,y)\sim\xi}\left[\mathbb{I}\left[x_j \neq y_j\right]\right]}_{=\mathbb{E}_{(x,y)\sim\xi}[d_H(x,y)]} \\
&= \mathscr{W}_1\left(\mu^{i\leftarrow+1}, \mu^{i\leftarrow-1}\right).
\end{aligned}$$

□

We now show that having a contractive coupling proof of fast mixing for a local Markov chain like Glauber dynamics implies $O(1)$-coupling independence.

Lemma 3.2 ([Liu21; Bla+22])**.** *Suppose the distribution μ satisfies* Dobrushin's uniqueness criterion*, i.e. $\|\mathscr{R}_\mu\|_{\ell_\infty\to\ell_\infty} \leq 1-\epsilon$ for some constant $\epsilon > 0$, where recall*

$$\mathscr{R}_\mu(i \to j) \stackrel{\mathsf{def}}{=} \max_{\tau:[n]\setminus\{i,j\}\to\{\pm1\}} d_{\mathsf{TV}}\left(\mu_j^{\tau,i\leftarrow+1}, \mu_j^{\tau,i\leftarrow-1}\right), \qquad \forall i \neq j,$$

and $\mathscr{R}_\mu(i \to i) = 0$ for all $i \in [n]$. Then μ is $\frac{1-\epsilon}{\epsilon}$-coupling independent.

Remark 1. Lemma 3.2 can be generalized considerably. In particular, one can replace Glauber dynamics with any Markov chain which changes $O(1)$-many coordinates in each step. One can also allow for *variable-length/multi-step* couplings of the chain, so long as Hamming distance contracts roughly by a constant factor every $O(n)$-steps. Finally, one can even allow weighted Hamming metrics. The main technique goes under the name *Stein's method for Markov chains*, and is a generically useful tool for bounding the transportation distance between distributions; see e.g. [BN19; RR19].

We before we get into the details of the proof, let us describe a generic strategy for bounding $\mathscr{W}_1(\nu,\pi)$; one can then of course replace ν,π with $\mu^{i\leftarrow+1}, \mu^{i\leftarrow-1}$. Fix a test function $f : \{\pm1\}^n \to \mathbb{R}$ which is 1-Lipschitz w.r.t. Hamming distance. If we can bound $|\mathbb{E}_\nu[f] - \mathbb{E}_\pi[f]|$ for all such f, then we get a bound on $\mathscr{W}_1(\nu,\pi)$ just by *Kantorovich–Rubinstein duality* for Wasserstein distance.

A natural strategy for constructing an "efficient" coupling/transport plan between ν,π is to use a Markov chain which simultaneously mixes rapidly, yet makes only "local moves". Let P_π be an ergodic Markov chain with stationary measure π (e.g. Glauber dynamics); we initialize it with ν. Then

$$\begin{aligned}
|\mathbb{E}_\nu[f] - \mathbb{E}_\pi[f]| &= \left|\sum_{t=0}^\infty \left(\nu^\top \mathsf{P}_\pi^t f - \nu^\top \mathsf{P}_\pi^{t+1} f\right)\right| && \text{(Telescoping)} \\
&\leq \sum_{t=0}^\infty \left|\nu^\top(\mathsf{Id} - \mathsf{P}_\pi)\mathsf{P}_\pi^t f\right| && \text{(Triangle Inequality)} \\
&\leq \mathop{\mathbb{E}}_{\substack{x\sim\nu \\ y\sim\mathsf{P}_\pi(x\to\cdot)}}\left[\sum_{t=0}^\infty \left|\left(\mathsf{P}_\pi^t f\right)(x) - \left(\mathsf{P}_\pi^t f\right)(y)\right|\right] && \text{(More Triangle Inequality)} \\
&= \mathop{\mathbb{E}}_{\substack{x\sim\nu \\ y\sim\mathsf{P}_\pi(x\to\cdot)}}\left[\sum_{t=0}^\infty \mathbb{E}_{(X_t,Y_t)}\left[|f(X_t) - f(Y_t)| \mid \substack{X_0=x \\ Y_0=y}\right]\right],
\end{aligned}$$

where $\{(X_t, Y_t)\}_{t=0}^{\infty}$ is any coupling of P_π. Intuitively, this seems very good if P_π is local because $y \sim \mathsf{P}_\pi(x \to \cdot)$ means $d_H(x, y) = d_H(X_0, Y_0)$ is small. If in addition P_π admits a coupling proof of rapid mixing (e.g. π satisfies Dobrushin's condition), then we also expect $\mathbb{E}_{(X_t,Y_t)}\left[|f(X_t) - f(Y_t)|\right]$ to quickly decay in t.

Unfortunately this isn't quite good enough because the rate of decay isn't fast enough. Indeed, for Glauber dynamics say, the best one can hope for is $\mathbb{E}_{(X_t,Y_t)}\left[|f(X_t) - f(Y_t)|\right] \le \left(1 - \frac{\epsilon}{n}\right)^t$ for some constant $0 < \epsilon < 1$. Hence, we would pay ≈ 1 for each t up to $O(n)$, leading to $|\mathbb{E}_\nu[f] - \mathbb{E}_\pi[f]| \le O(n)$. But we always trivially have $|\mathbb{E}_\nu[f] - \mathbb{E}_\pi[f]| \le n$ by 1-Lipschitzness of f, so it seems we have achieved very little.

The trick that will save us is to replace $\mathsf{Id} - \mathsf{P}_\pi$ in the first inequality step with $\mathsf{P}_\nu - \mathsf{P}_\pi$ for some other Markov chain P_ν whose stationary measure is ν. This trick will allow us to pick up a factor of $\max_x d_{\mathsf{TV}}\left(\mathsf{P}_\nu(x \to \cdot), \mathsf{P}_\pi(x \to \cdot)\right)$, which will be $O(1/n)$ for a good choice of P_ν rather than $O(1)$. Let us now formalize this strategy.

Proof of Lemma 3.2. For notational convenience, we write ν, π instead of $\mu^{i \leftarrow +1}, \mu^{i \leftarrow -1}$, which are supported on $\{\pm 1\}^{[n]\setminus\{i\}}$. Let $\mathsf{P}_\nu, \mathsf{P}_\pi$ denote Glauber dynamics w.r.t. ν, π, respectively. Then for every 1-Lipschitz test function f,

$$
\begin{aligned}
|\mathbb{E}_\nu[f] - \mathbb{E}_\pi[f]| &\le \sum_{t=0}^{\infty} \left|\nu^\top (\mathsf{P}_\nu - \mathsf{P}_\pi)\mathsf{P}_\pi^t f\right| && \text{(Triangle Inequality)} \\
&\le \mathbb{E}_{x\sim\nu}\left[\sum_{t=0}^{\infty}\left|\sum_y \left(\mathsf{P}_\nu(x \to y) - \mathsf{P}_\pi(x \to y)\right)\cdot\left(\mathsf{P}_\pi^t f\right)(y)\right|\right] \\
& && \text{(More Triangle Inequality)} \\
&\le \mathbb{E}_{x\sim\nu}\left[\sum_{t=0}^{\infty}\left|\sum_{y\neq x} \left(\mathsf{P}_\nu(x \to y) - \mathsf{P}_\pi(x \to y)\right)\cdot\left(\left(\mathsf{P}_\pi^t f\right)(y) - \left(\mathsf{P}_\pi^t f\right)(x)\right)\right|\right] \\
& && \text{(Using } \mathsf{P}(x \to x) = 1 - \textstyle\sum_{y\neq x}\mathsf{P}(x \to y)\text{)} \\
&\le \mathbb{E}_{x\sim\nu}\left[\sum_{y\neq x}\left|\mathsf{P}_\nu(x \to y) - \mathsf{P}_\pi(x \to y)\right| \cdot \sum_{t=0}^{\infty}\mathbb{E}_{(X_t,Y_t)}\left[|f(X_t) - f(Y_t)| \mid {\scriptstyle X_0 = x \atop Y_0 = y}\right]\right] \\
&\le \mathbb{E}_{x\sim\nu}\left[\sum_{y\neq x}\left|\mathsf{P}_\nu(x \to y) - \mathsf{P}_\pi(x \to y)\right| \cdot \sum_{t=0}^{\infty}\mathbb{E}_{(X_t,Y_t)}\left[d_H(X_t, Y_t) \mid {\scriptstyle X_0 = x \atop Y_0 = y}\right]\right], \\
& && \text{(1-Lipschitzness of } f\text{)}
\end{aligned}
$$

where $\{(X_t, Y_t)\}_{t=0}^{\infty}$ is any coupling of P_π. Now, let us recall that $\nu = \mu^{i \leftarrow +1}, \pi = \mu^{i \leftarrow -1}$. Since μ satisfies Dobrushin's condition, so do ν, π. Hence, we take $\{(X_t, Y_t)\}_{t=0}^{\infty}$ to be the one induced by one-step greedy path coupling, and so

$$
\mathbb{E}_{(X_t,Y_t)}\left[d_H(X_t, Y_t) \mid {\scriptstyle X_0 = x \atop Y_0 = y}\right] \le d_H(x, y) \cdot \sum_{t=0}^{\infty}\left(1 - \frac{\epsilon}{n-1}\right)^t = \frac{n-1}{\epsilon} \cdot d_H(x, y).
$$

Since we're looking at Glauber dynamics, only y satisfying $d_H(x, y) = 1$ will contribute to the sum over $y \neq x$. It follows that

$$
\begin{aligned}
|\mathbb{E}_\nu[f] - \mathbb{E}_\pi[f]| &\le \frac{n-1}{\epsilon} \cdot \max_{\sigma : [n]\setminus\{i\} \to \{\pm 1\}}\left\{\sum_{j\neq i}\left|\mathsf{P}_\nu\left(\sigma \to \sigma^{\oplus j}\right) - \mathsf{P}_\pi\left(\sigma \to \sigma^{\oplus j}\right)\right|\right\} \\
&\le \frac{1}{\epsilon} \cdot \max_\sigma\left\{\sum_{j\neq i}\left|\frac{\nu\left(\sigma^{\oplus j}\right)}{\nu(\sigma) + \nu\left(\sigma^{\oplus j}\right)} - \frac{\pi\left(\sigma^{\oplus j}\right)}{\pi(\sigma) + \pi\left(\sigma^{\oplus j}\right)}\right|\right\} \\
&\le \frac{1}{\epsilon} \cdot \sum_{j\neq i}\left|\mathscr{R}_\mu(i \to j)\right| && \text{(Using } \nu = \mu^{i \leftarrow +1}, \pi = \mu^{i \leftarrow -1}\text{)} \\
&\le \frac{1-\epsilon}{\epsilon}.
\end{aligned}
$$

□

References

[Ana+22] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. "Entropic Independence: Optimal Mixing of down-up Random Walks". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2022. Rome, Italy: Association for Computing Machinery, 2022, pp. 1418–1430 (cit. on p. 1).

[Ana+23] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. "Universality of Spectral Independence with Applications to Fast Mixing in Spin Glasses". In: *arXiv preprint arXiv:2307.10466* (2023) (cit. on p. 4).

[Bla+22] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. "On Mixing of Markov Chains: Coupling, Spectral Independence, and Entropy Factorization". In: *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2022, pp. 3670–3692 (cit. on p. 6).

[BN19] Guy Bresler and Dheeraj Nagaraj. "Stein's method for stationary distributions of Markov chains and application to Ising models". In: *The Annals of Applied Probability* 29.5 (2019) (cit. on p. 6).

[CE22] Yuansi Chen and Ronen Eldan. "Localization Schemes: A Framework for Proving Mixing Bounds for Markov Chains (extended abstract)". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2022, pp. 110–122. DOI: 10.1109/FOCS54457.2022.00018 (cit. on p. 1).

[CZ23] Xiaoyu Chen and Xinyuan Zhang. "A Near-Linear Time Sampler for the Ising Model with External Field". In: *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2023, pp. 4478–4503. DOI: 10.1137/1.9781611977554.ch170. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611977554.ch170 (cit. on p. 5).

[KKS21] Tali Kaufman, Rasmus Kyng, and Federico Soldá. "Scalar and Matrix Chernoff Bounds from ℓ_∞-Independence". In: *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2021, pp. 3732–3753 (cit. on p. 5).

[Liu21] Kuikui Liu. "From Coupling to Spectral Independence and Blackbox Comparison with the Down-Up Walk". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2021)*. Vol. 207. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 32:1–32:21 (cit. on p. 6).

[RR19] Gesine Reinert and Nathan Ross. "Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs". In: *The Annals of Applied Probability* 29 (5 2019) (cit. on p. 6).

[WJ08] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Vol. 1. Foundations and Trends in Machine Learning 1–2. Now Publishers Inc, 2008, pp. 1–305 (cit. on pp. 3, 4).

A Unfinished Proofs

Proof Sketch of Theorem 1.4. We use Lagrange multipliers. Viewing ν as a vector in $\mathbb{R}^\Omega$, consider the Lagrangian

$$\mathscr{L}_\mu(\nu, \boldsymbol{\theta}, \lambda) \stackrel{\text{def}}{=} \sum_{x\in\Omega} \nu(x) \log \frac{\nu(x)}{\mu(x)} + \sum_{i=1}^{d} \theta_i \cdot \left(m_i - \sum_{x\in\Omega} \nu(x)\varphi_i(x) \right) + \lambda \left(1 - \sum_{x\in\Omega} \nu(x) \right).$$

We have $\nabla_{\boldsymbol{\theta}} \mathscr{L}_\mu(\nu, \boldsymbol{\theta}, \lambda) = 0$ and $\partial_\lambda \mathscr{L}_\mu(\nu, \boldsymbol{\theta}, \lambda) = 0$ if and only if the constraints $\mathbb{E}_{x\sim\nu}[\boldsymbol{\varphi}(x)] = \boldsymbol{m}$ and $\sum_{x\in\Omega} \nu(x) = 1$ are satisfied. Furthermore,

$$\nabla_{\boldsymbol{\theta}} \mathscr{L}_\mu(\nu, \boldsymbol{\theta}, \lambda) = 1 + \log \frac{\nu}{\mu} - \langle \boldsymbol{\theta}, \boldsymbol{\varphi} \rangle - \lambda,$$

which yields zero if and only if $\nu(x) \propto \mu(x) \cdot \exp(\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x) \rangle)$ for all $x \in \Omega$. Here, the constant $e^{\lambda-1}$ gives the constant of proportionality, and just ensures that ν is normalized to $\sum_{x\in\Omega} \nu(x) = 1$. □

Proof of Theorem 1.5. Convexity of $f \mapsto \log \mathbb{E}_{x\sim\mu}\left[e^{f(x)}\right]$ follows by Cauchy–Schwarz:

$$\begin{aligned}\log \mathbb{E}_\mu\left[e^{(f+g)/2}\right] &= \log\left(\sum_{x\in\Omega}\sqrt{\mu(x)e^{f(x)}}\cdot\sqrt{\mu(x)e^{g(x)}}\right)\\ &\le \log\left(\sqrt{\sum_{x\in\Omega}\mu(x)e^{f(x)}}\cdot\sqrt{\sum_{x\in\Omega}\mu(x)e^{g(x)}}\right)\\ &= \frac{1}{2}\log\mathbb{E}_\mu\left[e^f\right]+\frac{1}{2}\log\mathbb{E}_\mu\left[e^g\right].\end{aligned}$$

We also get strict convexity (except when $f = g + c$ for a constant c) just by using the equality case of Cauchy–Schwarz. Strict convexity of $\nu \mapsto \mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu)$ just follows from strict convexity of the univariate function $x \mapsto x \log x$. Now that we have strict convexity, the remaining claims can be established by checking first-order stationarity conditions. □

B Unfinished Calculations with $\mathcal{F}_\mu(\cdot)$

In the proof of Theorem 1.1, we claimed that $\mathcal{F}_\mu(\boldsymbol{m})-\mathcal{F}_\mu(\boldsymbol{m}^*)-\langle\nabla\mathcal{F}_\mu(\boldsymbol{m}^*),\boldsymbol{m}-\boldsymbol{m}^*\rangle = \mathcal{F}_{\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m}^*)}\mu}(\boldsymbol{m})$. We justify this here. We first separate terms involving $\mathcal{L}^*_\mu(\cdot)$ and $\Phi_\mu(\cdot)$ in the left-hand side, which becomes

$$(1+\eta)\cdot\underbrace{\left[\mathcal{L}^*_\mu(\boldsymbol{m})-\mathcal{L}^*_\mu(\boldsymbol{m}^*)-\langle\nabla\mathcal{L}^*_\mu(\boldsymbol{m}^*),\boldsymbol{m}-\boldsymbol{m}^*\rangle\right]}_{(A)}-\underbrace{\left[\Phi_\mu(\boldsymbol{m})-\Phi_\mu(\boldsymbol{m}^*)-\langle\nabla\Phi_\mu(\boldsymbol{m}^*),\boldsymbol{m}-\boldsymbol{m}^*\rangle\right]}_{(B)}$$

We must show that $(A) = \mathcal{L}^*_{\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m}^*)}\mu}(\boldsymbol{m})$ and $(B) = \Phi_{\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m}^*)}\mu}(\boldsymbol{m})$. We do each in turn.

$$\begin{aligned}(A) &= \mathscr{D}_{\mathrm{KL}}\left(\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m})}\mu\,\|\,\mu\right)-\mathscr{D}_{\mathrm{KL}}\left(\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m}^*)}\mu\,\|\,\mu\right)-\langle\boldsymbol{\theta}^*(\boldsymbol{m}^*),\boldsymbol{m}-\boldsymbol{m}^*\rangle\\ &= \langle\boldsymbol{\theta}^*(\boldsymbol{m}),\boldsymbol{m}\rangle-\mathcal{L}_\mu(\boldsymbol{\theta}^*(\boldsymbol{m}))-\langle\boldsymbol{\theta}^*(\boldsymbol{m}^*),\boldsymbol{m}^*\rangle+\mathcal{L}_\mu(\boldsymbol{\theta}^*(\boldsymbol{m}^*))-\langle\boldsymbol{\theta}^*(\boldsymbol{m}^*),\boldsymbol{m}-\boldsymbol{m}^*\rangle\\ &= \langle\boldsymbol{\theta}^*(\boldsymbol{m})-\boldsymbol{\theta}^*(\boldsymbol{m}^*),\boldsymbol{m}\rangle-\log\frac{\mathbb{E}_{\sigma\sim\mu}\left[e^{\langle\boldsymbol{\theta}^*(\boldsymbol{m}),\sigma\rangle}\right]}{\mathbb{E}_{\sigma\sim\mu}\left[e^{\langle\boldsymbol{\theta}^*(\boldsymbol{m}^*),\sigma\rangle}\right]}\\ &= \langle\boldsymbol{\theta}^*(\boldsymbol{m})-\boldsymbol{\theta}^*(\boldsymbol{m}^*),\boldsymbol{m}\rangle-\mathcal{L}_{\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m}^*)}\mu}\left(\boldsymbol{\theta}^*(\boldsymbol{m})-\boldsymbol{\theta}^*(\boldsymbol{m}^*)\right)\\ &= \mathcal{L}^*_{\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m}^*)}\mu}(\boldsymbol{m}).\end{aligned}$$

For (B), it is more convenient to rewrite $\Phi_\mu(\boldsymbol{m})$ as

$$\Phi_\mu(\boldsymbol{m}) = \sum_{i=1}^n\left(\frac{1}{2}\log\left(1-m_i^2\right)+\frac{m_i}{2}\log\frac{1+m_i}{1-m_i}\right)-\sum_{i=1}^n\frac{m_i}{2}\log\frac{1+m_i(\mu)}{1-m_i(\mu)}-\underbrace{\frac{1}{2}\sum_{i=1}^n\log\left(1-m_i(\mu)^2\right)}_{\text{Independent of }\boldsymbol{m}}.$$

Then

$$\begin{aligned}(B) &= \sum_{i=1}^n\left(\frac{1}{2}\log\left(1-m_i^2\right)+\frac{m_i}{2}\log\frac{1+m_i}{1-m_i}\right)-\sum_{i=1}^n\left(\frac{1}{2}\log\left(1-(m_i^*)^2\right)+\frac{m_i^*}{2}\log\frac{1+m_i^*}{1-m_i^*}\right)\\ &\qquad+\left\langle\boldsymbol{m}-\boldsymbol{m}^*,\frac{1}{2}\log\frac{1-\boldsymbol{m}(\mu)}{1+\boldsymbol{m}(\mu)}-\nabla\Phi_\mu(\boldsymbol{m}^*)\right\rangle\\ &= \Phi_{\mathcal{T}_{\boldsymbol{\theta}^*(\boldsymbol{m}^*)}\mu}(\boldsymbol{m}).\end{aligned}$$