Lecture 3.

Today: Optimal recovery & statistical/computational gaps.

Recall: Last lecture, we discussed general strategies to design algorithms which perform recovery above the information theoretic threshold.

Ambitous goal: The models discussed so far are Bayesian models. Can one do Bayes optimal inference?

Today: Recent progress for $\mathbb{Z}_2$-synchronization problem.

Setup: $M = \frac{\lambda}{n}\sigma\sigma^T + W$. $\sigma \in \{\pm 1\}^n$.

NB: Sign symmetry $\Rightarrow$ σ can only be recovered upto sign.

Fix: Estimate the rank 1 matrix $\sigma\sigma^T$.

Loss function $L(X) = \frac{1}{n^2}\|X - \sigma\sigma^T\|_F^2$ $\Rightarrow$ Bayes optimal estimator

$$\hat{X}_{Bayes} = \mathbb{E}[\sigma\sigma^T / M].$$

The corresponding loss

$$MMSE(\lambda) = \frac{1}{n^2}\mathbb{E}\left[\|\hat{X}_{Bayes} - \sigma\sigma^T\|_F^2\right]$$

NB: $\hat{X}_{Bayes}$ involves a sum over the hypercube, & is thus computationally intractable!

Thm (Deshpande, Abbe, Montanari '16)

$$\lim_{n\to\infty} MMSE(\lambda) = \begin{cases} 1 & \text{if } \lambda \le 1 \\ 1 - q_*^2 & \text{if } \lambda > 1 \end{cases}$$

(q_* implicitly defined)

Q. Does there exist a computationally tractable estimator which achieves Bayes optimal performance?

Two provable strategies for $\mathbb{Z}_2$-synchronization.

(i) Approximate Message Passing

(ii) Variational inference based on the TAP free energy.

Approximate Message Passing (Montanari - Venkataramanan '19)

$M = \frac{\lambda}{n}\sigma\sigma^T + W$. Let φ_1 = top eigenvector of M. (with norm 1)

Alg: (i) Initialize: $x^0 = \sqrt{n}\,\varphi_1$.

(ii) Iteration: $x^{t+1} = M f_t(x^t) - b_t f_{t-1}(x^{t-1})$, $b_t = \frac{1}{n}\sum_{i=1}^{n} f_t'(x_i^t)$.

$\left[f_t(x) = (f_t(x_1), \dots, f_t(x_n))^T \in \mathbb{R}^n \right]$

Heuristics: First assume in addition to M, we have $y \sim N(\mu_0\sigma, \sigma_0^2 I)$. Can do Bayes optimal recovery of σ given y! $\to$ Does not utilize M at all!

Preliminary estimator $f_0(y)$.

Idea: $x^1 = M f_0(y) = \frac{\lambda}{n}\langle \sigma, f_0(y)\rangle \sigma + \underbrace{W f_0(y)}_{\sim N\left(0, \frac{\|f_0(y)\|^2}{n}\right)}$

$$\Rightarrow \quad x^1 \overset{d}{=} N(\mu_1\sigma, \sigma_1^2 I)$$

$$\mu_1 = \lambda\, \mathbb{E}\left[\Sigma_0 f_0(\mu_0\Sigma_0 + \sigma_0 G)\right]$$

$$\sigma_1^2 = \mathbb{E}\left[f_0(\mu_0\Sigma_0 + \sigma_0 G)^2\right]$$

$\Sigma_0 \sim \begin{cases} +1 & \text{wp } 1/2 \\ -1 & \end{cases}$ $\Big\}$ ind

$G \sim N(0,1)$

Idea: Keep repeating this algorithm!

__Issue__: For higher iterations, the output is no longer gaussian!

__Fix__: AMP!

__Thm__ [State Evolution] (Bayati-Montanari '11, ...)

Assume $\lambda > 1$ & wlog $\langle \sigma, \varphi_1 \rangle \geq 0$. Let $f_t : \mathbb{R} \to \mathbb{R}$ be Lipschitz $\forall t \geq 1$.

$(\mu_t, \sigma_t)_{t \geq 0}$ defined via

$$\mu_{t+1} = \lambda\, \mathbb{E}\left[\Sigma_0 f_t(\mu_t \Sigma_0 + \sigma_t G)\right], \quad \sigma_{t+1}^2 = \mathbb{E}\left[f_t(\mu_t \Sigma_0 + \sigma_t G)^2\right]$$

$$\mu_0 = \sqrt{1 - \frac{1}{\lambda^2}}, \quad \sigma_0 = 1/\lambda.$$

Now, if $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $|\psi(x) - \psi(y)| \leq C(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2$

$$\Rightarrow \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \psi(\Sigma_{0,i}, x_i^t) = \mathbb{E}\left[\psi(\Sigma_0, \mu_t \Sigma_0 + \sigma_t G)\right].$$

[True for a large class of test fns f_t]

__Idea__: Can we choose f_t judiciously to obtain Bayes optimal inference?

__Thm__ (Montanari-Venkataramanan '19) (informal)

If $\lambda > 1$, $\exists$ an optimal choice of f_t s.t. $\hat{x}^t(M) = \frac{f_t(x_t)}{\lambda}$ is achieves optimal recovery performance under the iterated limit $t \to \infty$ following $n \to \infty$.

Takeaway: Roughly Bayes optimal inference possible in some settings

Caveat: Iteration → need $t \to \infty$ before $n \to \infty$ (Li Wei- '22)

→ Bayes optimal AMP does optimal recovery!

(ii) Variational Inference & allied methods.

Background: Suppose we wish to estimate $\mathbb{E}[\sigma\sigma^T | M]$.

$\Leftrightarrow \forall\, i \neq j$, estimate $\mathbb{E}[\sigma_i \sigma_j | M]$ (low dimensional marginals of the posterior dist.

The full posterior is intractable...

Idea: Maybe to approximate low-dim marginals, we can approximate the posterior $p(\sigma|M)$ by a simpler distribution, eg product dist?

NMF inference : $\hat{q} = \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\, D_{KL}(q(\sigma) \| p(\sigma|M))$

NMF $\leftrightarrow$ $\mathcal{Q}$ – product distribution $\prod_{i=1}^{n} q_i$.

Given $q = \prod_{i=1}^{n} q_i$, $m_i = \sum_{\sigma_i \in \{\pm 1\}} \sigma_i q_i(\sigma_i)$,

$$D_{KL}(q \| p(\sigma|M)) = \mathcal{F}(m) + \text{const.}$$ (Ghorbani et. al. '18)

$$\mathcal{F}(m) = -\frac{\lambda}{2}\langle m, M_0 m\rangle - \sum_{i=1}^{n} h(m_i)$$

$$M_0 = M - \operatorname{diag}(M).$$

$$h(m_i) = -\frac{(1+m_i)}{2}\log\frac{1+m_i}{2} - \frac{(1-m_i)}{2}\log\frac{1-m_i}{2}.$$

Local instability for $m_* = 0$.

$$\nabla \mathcal{F}(m)\Big|_{m=m_*} = 0, \quad \nabla^2 \mathcal{F}(m)\Big|_{m=m_*} = -\lambda M_0 + I.$$

$$\lambda_{\min}\left(\nabla^2 \mathcal{F}\Big|_{m_*}\right) \longrightarrow \begin{cases} 1-2\lambda & \lambda \le 1 \\ -\lambda^2 & \lambda > 1. \end{cases}$$

$\Rightarrow$ m_* is local minimum for $\lambda < 1/2$, saddle point for $\lambda \in [1/2, 1]$.

Q. Can we fix this?

Alt conjecture:
$$\mathcal{F}_{TAP}(m) = -\frac{\lambda}{2n}\langle m, Mm\rangle - \frac{1}{n}\sum_{i=1}^{n} h(m_i) - \frac{\lambda^2}{4}\left[1 - Q(m)\right]^2$$

$$Q(m) = \frac{\|m\|_2^2}{n}$$

Thm: (Fan, Mei, Montanari '20, Celentano-Fan-Mei '21)
(informal)

Fix $\lambda > 1$. (i) $\exists$ a local minimizer m_* s.t. $\frac{1}{n}\|m_* m_*^T - \hat{X}_{Bayes}\|_F^2 \to 0$.

(ii) $\mathcal{F}_{TAP}$ is strongly convex in a $\sqrt{\varepsilon n}$-nbhd of m_*.

(iii) Spectral initialization + AMP fixed rounds + mirror descent on $\mathcal{F}_{TAP}$ $\to$ linear convergence to m_*.

Statistical - Computational gaps.

So far: If detection/weak recovery possible, also possible using computationally efficient algorithms.

Ex Community detection with k-groups.

$\sigma_i \overset{iid}{\sim} \mathcal{U}(\{1, 2, \ldots, k\})$ $\quad \mathbb{P}[\{i,j\} \in E \mid \sigma] = \begin{cases} a/n & \text{if } \sigma_i = \sigma_j \\ b/n & \text{o.w.} \end{cases}$

$$d = \frac{a + (r-1)b}{r}.$$

Q. When is detection/weak recovery possible?

Fact: All existing polynomial time algorithms for detection/weak recovery require $\lambda := \dfrac{a-b}{\sqrt{r(a+(r-1)b)}} > 1.$

(Kesten-Stigum threshold)

Abbe-Sandon $\rightarrow$ If exponential computational budget available, can do detection/weak recovery for $\lambda < 1$!

Coja oghlan et. al $\rightarrow$ Exact IT threshold for $b > a$.

Conjecture: No polytime algorithm can go below λ_{KS}!