STAT260 Mean Field Asymptotics in Statistical Learning Lecture 19 - 04/05/2021

Lecture 19: Approximate message passing algorithms

Lecturer: Song Mei *Scriber: Feynman Liang* *Proof reader: Alexander Tsigler*

1 Algorithm for Gibbs mean (Statistical estimators)

Consider a configuration space $\Omega \subset \mathbb{R}^d$ with base measure $\nu_0 \in \mathcal{P}(\mathbb{R}^d)$, recall a Gibbs distribution

$$P_\beta(d\boldsymbol{\sigma}) \propto e^{-\beta H(\boldsymbol{\sigma})}\nu_0(d\boldsymbol{\sigma})$$

at inverse temperature β is determined by a (random) Hamiltonian $H : \Omega \to \mathbb{R}$. We are interested in efficiently approximating ensemble averages $\langle \sigma \rangle_\beta \in \mathbb{R}^d$.

Example 1 (Bayes linear model, Bayes estimator, LASSO)**.** Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ with $x_{0i} \sim_{i.i.d.} \mathbb{P}_0$, $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{w}$, $w_i \sim_{i.i.d.} N(0, \sigma^2)$.

The posterior mean estimator is

$$\hat{\boldsymbol{x}}_{Bayes} = \int_{\mathbb{R}^d} \boldsymbol{x} p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{A})\mathrm{d}\boldsymbol{x} \in \mathbb{R}^d,$$
$$p(\boldsymbol{x} \mid \boldsymbol{A}, \boldsymbol{y}) \propto e^{-\frac{\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2}{2\sigma^2}} \prod_{i=1}^d \mathbb{P}_0(x_i).$$

The LASSO estimator

$$\hat{\boldsymbol{x}}_{LASSO} = \arg\min_{\boldsymbol{x}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\lambda}{n}\|\boldsymbol{x}\|_1 = \lim_{\beta\to\infty} \int_{\mathbb{R}^d} \boldsymbol{x} p_\beta(\boldsymbol{x} \mid \boldsymbol{A}, \boldsymbol{y})\mathrm{d}\boldsymbol{x},$$
$$p_\beta(\boldsymbol{x} \mid \boldsymbol{A}, \boldsymbol{y}) \propto \exp\left\{-\beta\left[\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\lambda}{n}\|\boldsymbol{x}\|_1\right]\right\}.$$

Remark 1. *We have seen many observables with concentrated ensemble averages, e.g.*

$$O(\boldsymbol{x}) = \frac{1}{d}\sum_{i=1}^d x_i^2 \implies \langle O \rangle_\beta \approx \mathbb{E}\langle O \rangle_\beta,$$

where the randomness comes from the Hamiltonian H (i.e. $\boldsymbol{A}$, $\boldsymbol{x}_0$, and $\boldsymbol{w}$ for the Bayes linear model). If $\langle O \rangle_\beta$ concentrates, it can be approximated independent of any specific realization of $\boldsymbol{A}$, $\boldsymbol{x}_0$, and $\boldsymbol{w}$ hence its limiting value depends on the distribution of the random Hamiltonian H.

However, ensemble averages of coordinates $O(\boldsymbol{x}) = x_i$ typically do not concentrate (i.e. $\langle x_i \rangle_\beta \not\approx \mathbb{E}\langle x_i \rangle_\beta$). Hence, approximating ensemble averages of coordinates $\langle x_i \rangle_\beta$ depends on a specific instance (i.e. realizations of $\boldsymbol{A}$, $\boldsymbol{x}_0$, $\boldsymbol{w}$) of the Hamiltonian H.

The approximate message passing (AMP) algorithm is used to calculate $\hat{\boldsymbol{x}} = \langle \boldsymbol{\sigma} \rangle_\beta$.

2 ISTA and FISTA for LASSO

The convex optimization problem defining LASSO has structure

$$\hat{\boldsymbol{x}}_{LASSO} = \arg\min_{\boldsymbol{x}} \underbrace{\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2}_{\text{convex differentiable } f(\boldsymbol{x})} + \underbrace{\frac{\lambda}{n}\|\boldsymbol{x}\|_1}_{\text{convex separable } g(\boldsymbol{x})}.$$

An algorithm for solving this is proximal gradient descent (PGD) / iterative thresholding (ISTA) algorithm, defined by iterates

$$\boldsymbol{x}^{k+1} = \arg\min_{\boldsymbol{x}} \left[\frac{1}{2\zeta_k} \|\boldsymbol{x} - \underbrace{(\boldsymbol{x}^k - \zeta_k \nabla f(\boldsymbol{x}^k))}_{=:\boldsymbol{x}^{k+1/2}}\|_2^2 + g(\boldsymbol{x}) \right].$$

This is also called *proximal gradient descent* (PGD), and has closed-form solution when g is separable. For LASSO,

$$\begin{aligned} \boldsymbol{x}^{k+1} &= \eta(\boldsymbol{x}^k - \zeta_k \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^k - \boldsymbol{y}); \lambda\zeta_k), \qquad \boldsymbol{x}^1 = 0, \\ \eta(x;\theta) &= (|x| - \theta) \cdot 1\{|x| > \theta\}. \end{aligned}$$

Theorem 2. *Suppose $f \in C^2(\mathbb{R}^d)$ convex, $\sup_{\boldsymbol{x}} \|\nabla^2 f(\boldsymbol{x})\|_{op} \leq \beta$, $g \in C(\mathbb{R}^d)$ convex, $\mathcal{C}(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$, $\arg\min_{\boldsymbol{x}} \mathcal{C}(\boldsymbol{x}) \neq \emptyset$. Then taking $\zeta_k = \frac{1}{\beta}$, the kth PGD iterate $\boldsymbol{x}^k$ has cost-function guarantee*

$$\mathcal{C}(\boldsymbol{x}^k) - \min_{\boldsymbol{x}} \mathcal{C}(\boldsymbol{x}) \leq \frac{\beta\|\boldsymbol{x}' - \boldsymbol{x}_k\|_2^2}{2k} = O\left(\frac{1}{k}\right).$$

Proof reference. Fully deterministic, based on Jensen's inequality and algebra. See [Beck and Teboulle, 2009]
. ∎

An accelerated variant called APGD / fast iterative soft-thresholding (FISTA) uses a momentum sequence $\mu_1 = 0$, $\mu_k = \frac{1+\sqrt{1+4\mu_{k-1}^2}}{2}$, $\gamma_k = \frac{1-\mu_k}{\mu_{k+1}}$ and defines iterates

$$\begin{aligned} \boldsymbol{\nu}^{k+1} &= \arg\min_{\boldsymbol{x}} \left[\frac{\beta}{2} \|\boldsymbol{x} - (\boldsymbol{x}^k - \frac{1}{\beta}\nabla f(\boldsymbol{x}^k))\|_2^2 + g(\boldsymbol{x}) \right], \\ \boldsymbol{x}^{k+1} &= (1-\gamma_k)\boldsymbol{\nu}^{k+1} + \gamma_k \boldsymbol{\nu}^k. \end{aligned}$$

Theorem 3. *Under the same assumptions, the iterates of APGD satisfy*

$$\mathcal{C}(\boldsymbol{x}^k) - \min_{\boldsymbol{x}} \mathcal{C}(\boldsymbol{x}) \leq \frac{2\beta\|\boldsymbol{x}' - \boldsymbol{x}_k\|_2^2}{k^2} = O\left(\frac{1}{k^2}\right).$$

3 The AMP algorithm for LASSO

ISTA with step size 1 has updates

$$\begin{aligned} \boldsymbol{x}^{k+1} &= \eta(\boldsymbol{x}^k + \boldsymbol{A}^\top \boldsymbol{z}^k; \theta_k), \\ \boldsymbol{z}^k &= \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^k, \end{aligned}$$

where $\theta_k = \lambda$.

Compare this to AMP (will show how to derive in later lectures), which has updates

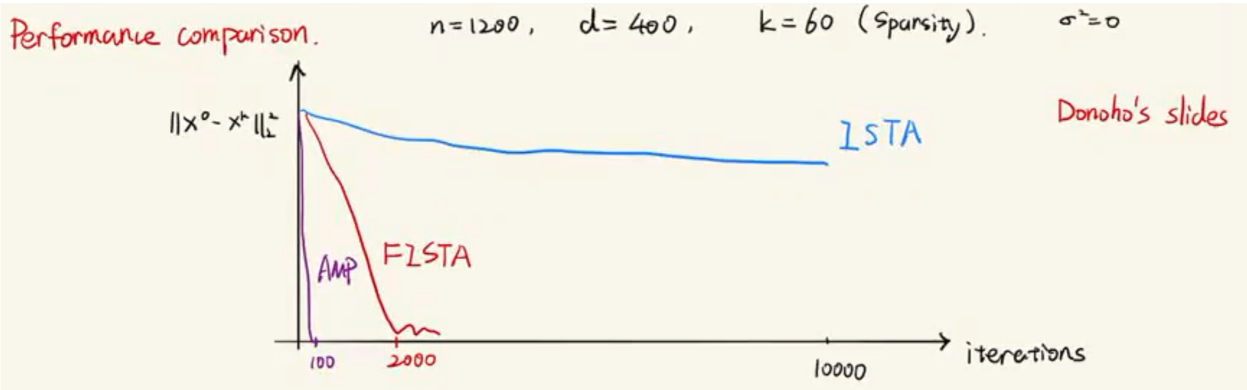
$$\begin{aligned} \boldsymbol{x}^{k+1} &= \eta(\boldsymbol{x}^k + \boldsymbol{A}^\top \boldsymbol{z}^k; \theta_k), \\ \boldsymbol{z}^k &= \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^k + \underbrace{\omega_k \boldsymbol{z}^{k-1}}_{\text{Onsager correction term}}, \end{aligned}$$

where $\omega_k = \frac{1}{d}\sum_{i=1}^d \eta'(\bar{x}_i^k; \theta_{k-1})$, where $\bar{\boldsymbol{x}}^k = \boldsymbol{x}^k + \boldsymbol{A}^\top \boldsymbol{z}^{k-1}$, and each θ_k is a suitably chosen scalar.

Remark 4. *In practice it is not suggested to use AMP to solve LASSO. This is because*

- *AMP is not monotonically decreasing in $\mathcal{C}$; not a deterministic proof of convergence.*
- *The convergence analysis of AMP depends on assumptions on $\boldsymbol{A}$ and $\boldsymbol{y}$.*

However, when distribution assumptions on $\boldsymbol{A}$ and $\boldsymbol{y}$ are satisfied typically AMP is faster than ISTA/FISTA.



Assumptions required for AMP:

1. $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with entries $A_{ij} \sim_{i.i.d.} N(0, 1/n)$.
2. $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\frac{1}{d}\sum_{i=1}^d \delta_{x_{0,i}} \Rightarrow \mathbb{P}_{\boldsymbol{x}_0}$, $\frac{1}{d}\sum_{i=1}^d x_{0,i}^2 \to \mathbb{E}_{\boldsymbol{x}_0}[x_{0,i}^2]$.
3. $\boldsymbol{w} \in \mathbb{R}^n$, $\frac{1}{n}\sum_{i=1}^n \delta_{w_i} \Rightarrow \mathbb{P}_w$, $\frac{1}{n}\sum_{i=1}^d w_i^2 \to \mathbb{E}_w[w_i^2]$.
4. $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{w} \in \mathbb{R}^n$.
5. $n/d \to \delta$.

4 Theoretical analysis of AMP

To understand and analyze AMP, it is helpful to consider a state evolution (SE) characterization

$$\tau_{k+1}^2 = F(\tau_k^2, \theta_k),$$

where

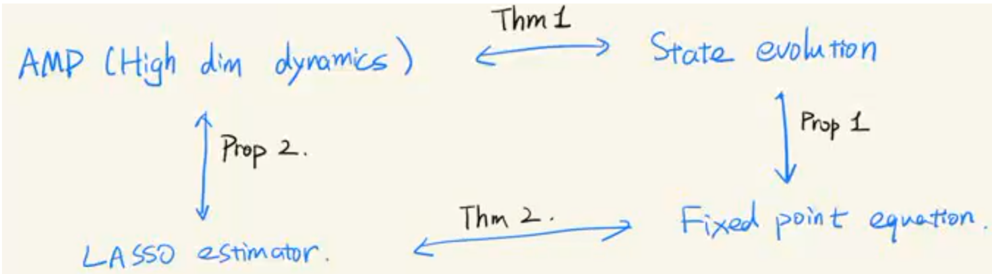
$$\begin{aligned} &F(\tau^2, \theta) := \sigma^2 + \frac{1}{\delta}\mathbb{E}[(\eta(X_0 + \tau G; \theta) - X_0)^2], \\ &(X_0, G) \sim \mathbb{P}_{X_0} \times \mathcal{N}(0, 1). \end{aligned}$$

AMP and SE are connected by the following key result:

Theorem 5 ([Bayati and Montanari, 2011])**.** *Let assumptions 1–5 hold. For any test pseudo-Lipschitz test function $\psi : \mathbb{R}^2 \to \mathbb{R}$ (i.e. $|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \le K\|\boldsymbol{x} - \boldsymbol{y}\|_2(1 + \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2)$), almost surely*

$$\lim_{\substack{d\to\infty \\ n/d\to\delta}} \frac{1}{d}\sum_{i=1}^d \psi(x_i^{k+1}, x_{0,i}) = \mathbb{E}[\psi(\eta(X_0 + \tau_k G, \theta_k), X_0)].$$

Remark 6. *This result says* $\delta(\tau_{k+1}^2 - \sigma^2) = \lim_{\substack{d\to\infty \\ n/d\to\delta}} \|\boldsymbol{x}^k - \boldsymbol{x}_0\|_2^2/d$*. So, to analyze the asymptotic behavior of (high dimensional) AMP on* $\boldsymbol{x}$*, it suffices to analyze the (low dimensional) behavior of SE. We will later see that SE converges to the solution of a fixed point, and another later result will show AMP converges to LASSO, yielding a chain of relations:*



Let (τ_*, α_*) be a proper solution of the fixed-point equations

$$\tau^2 = \sigma^2 + \delta^{-1}\mathbb{E}[(\eta(X_0 + \tau G; \alpha\tau) - X_0)^2],$$
$$\lambda = \alpha\tau \left(1 - \delta^{-1}\mathbb{E}[\eta'(X_0 + \tau G; \alpha\tau)]\right)^2,$$

with (uniqueness-enforcing) constraint $\delta \geq \delta_*(\sigma^2, \lambda)$.

Proposition 7. *With* $\theta_k = \alpha_*\tau_k$*, the state evolution* $\{\tau_k\}_{k\geq 1}$ *of* $\tau_{k+1}^2 = F(\tau_k^2, \alpha_*\tau_k)$ *converges (exponentially fast) to* τ_*^2*.*

Proposition 8. *Let* $\hat{\boldsymbol{x}}(\lambda) = \arg\min_{\boldsymbol{x}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\lambda}{n}\|\boldsymbol{x}\|_1$*,* $\{\boldsymbol{x}^k\}_{k\geq 1}$ *AMP iterates. Then*

$$\lim_{k\to\infty}\lim_{n\to\infty} \|\hat{\boldsymbol{x}}(\lambda) - \boldsymbol{x}^k\|_2^2/d = 0.$$

Proof strategy. Show $\exists \boldsymbol{v}_k \in \partial\mathcal{C}(\boldsymbol{x})$ (subgradients, since $\mathcal{C}$ not smooth) such that $\lim_{d\to\infty}\|\boldsymbol{v}_k\|_2^2/d = 0$ (depends on previous proposition). Then use convexity of $\mathcal{C}(\boldsymbol{x})$ to argue sequence converges to minimizer of $\mathcal{C}(\boldsymbol{x})$. □

Remark 9. *One can get exponential convergence in distance in the following sense: for any* k

$$\lim_{n\to\infty} \|\hat{\boldsymbol{x}}(\lambda) - \boldsymbol{x}^k\|_2^2/d \leq e^{-ck} \lim_{n\to\infty} \|\hat{\boldsymbol{x}}(\lambda) - \boldsymbol{x}^0\|_2^2/d.$$

Note this is differs from standard exponential convergence, which requires:

$$\|\hat{\boldsymbol{x}}(\lambda) - \boldsymbol{x}^k\|_2^2/d \leq e^{-ck}\|\hat{\boldsymbol{x}}(\lambda) - \boldsymbol{x}^0\|_2^2/d.$$

The connection between LASSO and the fixed point of SE closes the loop:

Theorem 10 ([Bayati and Montanari, 2011])**.** *Let assumptions 1 through 5 hold. For any test pseudo-Lipschitz test function* $\psi : \mathbb{R}^2 \to \mathbb{R}$ *(i.e.* $|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \leq K\|\boldsymbol{x} - \boldsymbol{y}\|_2(1 + \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2)$*), almost surely*

$$\lim_{\substack{d\to\infty \\ n/d\to\delta}} \frac{1}{d}\sum_{i=1}^{d} \psi(\hat{x}_i(\lambda), x_{0,i}) = \mathbb{E}[\psi(\eta(X_0 + \tau_* G; \alpha_*\tau_*), X_0)].$$

References

[Bayati and Montanari, 2011] Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.

[Beck and Teboulle, 2009] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202.