Mean Field Asymptotics in Statistical Learning.

Apr 5

Lecture 19. Approximate message passing algorithms.

① Algorithm for Gibbs mean (Statistical estimators).

Configuration space: $\Omega \subseteq \mathbb{R}^d$. Measure $\nu_0 \in \mathcal{P}(\mathbb{R}^d)$.

Random Hamiltonian: $H: \Omega \to \mathbb{R}$.

Gibbs distribution: $P_\beta(d\sigma) \propto \exp\{-\beta H(\sigma)\}\nu_0(d\sigma)$.

Goal: (approximately) calculate $\langle \sigma \rangle_\beta \in \mathbb{R}^d$ using efficient algorithm.

Example: Bayes linear model, Bayes estimator, and LASSO estimator.

$x_0 \in \mathbb{R}^d$, $x_{0i} \sim$ iid P_0. $A \in \mathbb{R}^{n \times d}$.

$y = Ax_0 + w$, $w_i \sim$ iid $N(0, \sigma^2)$.

Posterior mean estimator

$$\hat{x}_{Bayes} = \int_{\mathbb{R}^d} x \, P(x|y,A)\,dx \in \mathbb{R}^d$$

$$P(x|A,y) \propto \exp\left\{-\frac{\|y-Ax\|_2^2}{2\sigma^2}\right\} \prod_{i=1}^d P_0(x_i)$$

LASSO estimator:

$$\hat{x}_{LASSO} = \arg\min_x \left\{\frac{1}{2n}\|y-Ax\|_2^2 + \frac{\lambda}{n}\|x\|_1\right\} = \lim_{\beta\to\infty} \int_{\mathbb{R}^d} x \, P_\beta(x|y,A)\,dx$$

$$P_\beta(x|A,y) \propto \exp\left\{-\beta\left[\frac{1}{2n}\|y-Ax\|_2^2 + \frac{\lambda}{n}\|x\|_1\right]\right\}.$$

Remark: We have seen many concentrated observables,

e.g. $O(x) = \frac{1}{d}\sum_{i=1}^d x_i^2 \Rightarrow \langle O \rangle_\beta \approx \mathbb{E}[\langle O \rangle_\beta]$.

Randomness come from H, i.e. A, x_0, w.

So, approximating $\langle O \rangle_\beta$ can be independent of specific instance of A, x_0, w. Limiting value depends on their distribution instead of realization.

However, EA of $O(x) = x_i$ typically doesn't concentrate.

$$\langle x_i \rangle_\beta \not\approx \mathbb{E}[\langle x_i \rangle_\beta].$$

$\Rightarrow$ Approximating $\langle x_i \rangle_\beta$ needs to use the specific instance H.

AMP is an algorithm used to calculate $\hat{x} = \langle \sigma \rangle_\beta$.

② ISTA and FISTA for LASSO.

convex differentiable $f(x)$ — convex Separable $g(x)$.

$$\hat{x}_{LASSO} = \arg\min_x \mathcal{L}(x) = \arg\min_x \frac{1}{2n}\|y - Ax\|_2^2 + \frac{\lambda}{n}\|x\|_1.$$

Convex optimization problem.

PGD/ISTA (Proximal gradient descent ; Iterative thresholding algorithm)

$x^{k+\frac{1}{2}} = x^k - s_k \nabla f(x^k)$

$$x^{k+1} = \arg\min_x \left[\frac{1}{2s_k}\|x - (x^k - s_k \nabla f(x^k))\|_2^2 + g(x)\right] \quad \text{PGD}$$

Proximal : $x^{k+1} = \arg\min_x \left[\frac{1}{2s_k}\|x - x^{k+\frac{1}{2}}\|_2^2 + g(x)\right]$

$$\Rightarrow \quad x^{k+1} = \eta(x^k - s_k A^T(Ax^k - y); \lambda s_k) \qquad x^1 = 0.$$

$$\eta(x;\theta) = (|x| - \theta)\cdot 1\{|x| > \theta\}.$$

[Beck, Teboulle, 2008]

Thm: Suppose $f \in C^2(\mathbb{R})$, $\sup_x \|\nabla^2 f(x)\|_{op} \le \beta$, convex.

$g \in C(\mathbb{R})$, convex.

$\ell(x) = f(x) + g(x)$. $\arg\min_x \ell(x) \neq \phi$.

Then $s_k = \frac{1}{\beta}$, x^k : k'th PGD iteration.

$$\Rightarrow \quad \ell(x^k) - \min_x \ell(x) \le \frac{\beta\|x^1 - x_*\|_2^2}{2k} = O\left(\frac{1}{k}\right)$$

Proof is fully deterministic, based on Jenssen's inequality and simple algebra

APGD/FISTA. $\mu_1 = 0$. $\mu_k = \frac{1+\sqrt{1+4\mu_{k-1}^2}}{2}$. $r_k = \frac{1-\mu_k}{\mu_{k+1}}$.

$$\begin{cases} v^{k+1} = \arg\min_x \left\{\frac{\beta}{2}\|x - (x^k - \frac{1}{\beta}\nabla f(x^k))\|_2^2 + g(x)\right\} \\ x^{k+1} = (1-r_k)v^{k+1} + r_k v^k. \end{cases}$$

$$x^{k+1} = (1-r_k)\,\eta\left(x^k - \tfrac{1}{\beta}A^T(Ax^k - y); \tfrac{\lambda}{\beta}\right) + r_k\,\eta\left(x^{k-1} - \tfrac{1}{\beta}A^T(Ax^{k-1} - y); \tfrac{\lambda}{\beta}\right).$$

Thm: Under the same assumptions.

Let x^k : k'th iteration of APGD. then

$$\ell(x^k) - \min_x \ell(x) \le \frac{2\beta\|x^1 - x_*\|_2^2}{k^2} = O\left(\frac{1}{k^2}\right).$$

Remark:

③ The approximate algorithm passing algorithm for LASSO.

ISTA w/ step size 1.

$\eta'(x;\theta) = \partial_x \eta(x;\theta)$.

$$\begin{cases} x^{k+1} = \eta(x^k + A^T z^k ; \theta_k) \\ z^k = y - Ax^k \end{cases} \qquad \theta_k = \lambda.$$

AMP: (Will show how this can be derived in later lectures)

$$\begin{cases} x^{k+1} = \eta(x^k + A^T z^k ; \theta_k) \\ z^k = y - Ax^k + w_k z^{k-1} \end{cases} \qquad \theta_k, w_k \text{ scalers.}$$

Onsager correction term. (Intuition will be given later)

$$w_k = \frac{1}{\delta} \sum_{i=1}^{d} \eta'(x^{k-1} + A^T z^{k-1} ; \theta_{k-1});$$

$\eta'(x^{k-1} + A^T z^{k-1} ; \theta_{k-1}) \in \mathbb{R}^d$

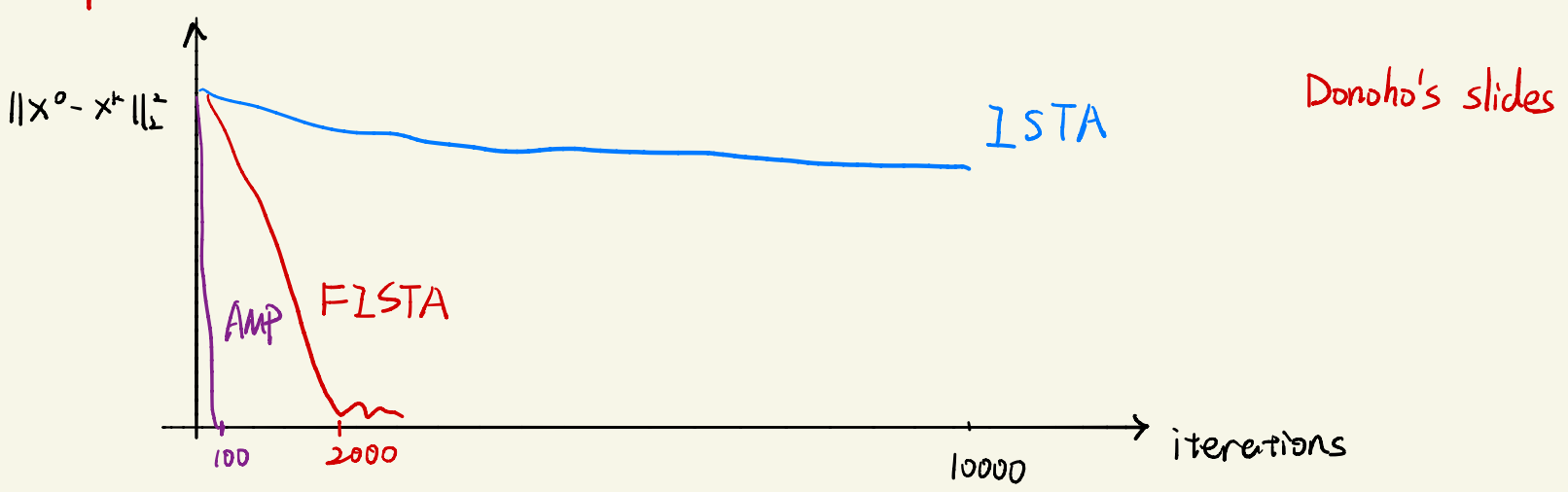
θ_k: needs to be properly chosen. Subtle. Now just assume arbitrary.

Remark: ⊗ AMP is not descending. Not a deterministic proof of convergence.

⊗ Convergence analysis of AMP depend on assuptions of A, y.

Performance comparison. $n = 1200$, $d = 400$, $k = 60$ (Sparsity). $\sigma^2 = 0$

Donoho's slides



Assuptions:

a) $A \in \mathbb{R}^{n \times d}$. $A_{ij} \sim_{iid} N(0, \frac{1}{n})$.

b) $x_0 \in \mathbb{R}^d$. $\frac{1}{d}\sum_{i=1}^{d} \delta_{x_{0,i}} \xrightarrow{\text{weak}} \mathbb{P}_{x_0}$ $\frac{1}{d}\sum_{i=1}^{d} x_{0,i}^2 \to \mathbb{E}_{x_0}(x_0^2]$

c) $w \in \mathbb{R}^n$. $\frac{1}{n}\sum_{i=1}^{n} \delta_{w_i} \xrightarrow{\text{weak}} \mathbb{P}_w$, $\frac{1}{n}\sum_{i=1}^{n} w_i^2 \to \mathbb{E}_w[w^2] = \sigma^2$.

d) $y = Ax_0 + w \in \mathbb{R}^n$.

e) $n/d \to \delta$.

④ Theoretical analysis of AMP.

State evolution characterization:

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

where $F(\tau^2, \theta) \equiv \sigma^2 + \frac{1}{\delta}\mathbb{E}[(\eta(X_0 + \tau G; \theta) - X_0)^2]$.

$(X_0, G) \sim \mathbb{P}_{X_0} \times N(0,1)$.

Thm 1 ([Bayati, Montanari, 2011]).

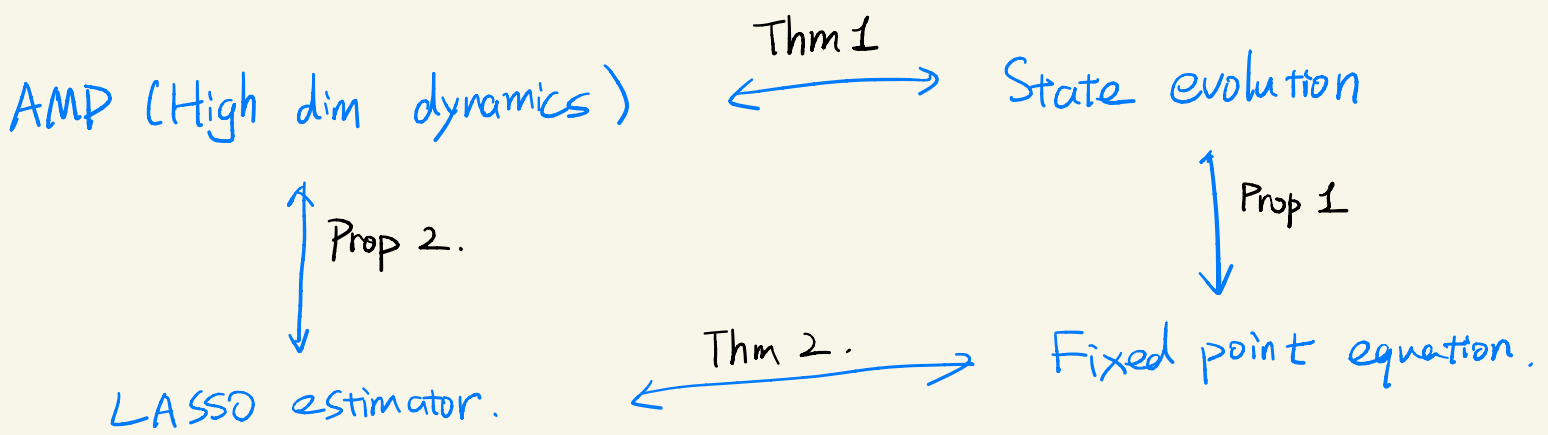
Let assumptions a) — e) hold. Let $\psi: \mathbb{R}^2 \to \mathbb{R}$.

be Pseudo Lipschitz. i.e. $|\psi(x) - \psi(y)| \le K \cdot \|x-y\|_2 \cdot (1 + \|x\|_2 + \|y\|_2)$

Then almost surely.

$$\lim_{\substack{d\to\infty \\ n/d \to \delta}} \frac{1}{d}\sum_{i=1}^{d} \psi(x_i^{k+1}, x_{0,i}) = \mathbb{E}[\psi(\eta(x_0 + \tau_k G; \theta_k), X_0)].$$

Rmk: $\delta(\tau_{k+1}^2 - \sigma^2) = \lim_{\substack{d\to\infty \\ n/d\to\delta}} \|x^k - x_0\|_2^2/d$ $\qquad \psi(x, x_0) = (x - x_0)^2$



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

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

$$\delta \gtrless \delta_*(\sigma^2, \lambda)$$

Prop 1: Take $\theta_k = \alpha_* \tau_k$, the state evolution $\{\tau_k^2\}_{k\ge 1}$.

$\tau_{k+1}^2 = F(\tau_k^2, \alpha_* \tau_k)$ converge to τ_*^2.

The convergence is exponentially fast.

Prop 2: Let $\hat{x}(\lambda) \equiv \arg\min_x \underbrace{\frac{1}{2n}\|y - Ax\|_2^2 + \frac{\lambda}{n}\|x\|_1}_{e(x)}$

$\{x^k\}_{k\geq 1}$ AMP iterates. Then

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

Proof strategy: ⊗ Show $\exists\, v_k \in \partial e(x^k)$ (Thm 1).

$$\lim_{k\to\infty} \lim_{d\to\infty} \|v_k\|_2^2/d = 0.$$

⊗ Convex property of $e(x)$.

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

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

Thm 2: ([Bayati, Montanari, 2011]).

Let assumptions a) – e) holds, Let $\psi: \mathbb{R}^2 \to \mathbb{R}$ be any pseudo-Lipschitz function. i.e. $|\psi(x) - \psi(y)| \leq K \cdot \|x-y\|_2 \cdot (1 + \|x\|_2 + \|y\|_2)$.

Then almost surely

$$\lim_{\substack{d\to\infty \\ n/d\to\delta}} \frac{1}{d}\sum_{i=1}^{d} \psi(\hat{x}(\lambda), x_{0,i}) = \mathbb{E}\left[\psi(\eta(x_0 + \tau_\delta G; \alpha_\delta \tau_\delta); X_0)\right].$$

Remark: Similar analysis and results for a wide range of problem.

Z₂ synchro. Generalized linear models.

Plans for remaining lectures: