

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: $H: \Omega \rightarrow \mathbb{R}$.

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.

$x_0 \in \mathbb{R}^d$, $x_{gi} \sim \text{iid } \mathbb{P}_0$, $A \in \mathbb{R}^{n \times d}$.

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

Posterior mean estimator

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

$$P(x|A, y) \propto \exp\left\{-\frac{\|y - Ax\|_2^2}{2\sigma^2}\right\} \prod_{i=1}^d \mathbb{P}_0(x_i)$$

LASSO estimator:

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

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

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, x₀, 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.

$$\langle x_i \rangle_\beta \neq \mathbb{E}[\langle x_i \rangle_\beta].$$

$\Rightarrow$ 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

convex
separable.

$f(x)$
//

$g(x)$
//

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

Convex optimization problem.

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

$$x^{k+1} = \underset{x}{\operatorname{argmin}} \left[\frac{1}{2\zeta_k} \|x - (x^k - \zeta_k \nabla f(x^k))\|_2^2 + g(x) \right] \quad \text{PGD}$$

Check: stationery condition gives $0 \in \partial \mathcal{L}(x_*)$.

$$\Rightarrow x^{k+1} = \eta(x^k - \zeta_k A^T(Ax^k - y); \lambda \zeta_k), \quad x^1 = \text{arbitrary}$$

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

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

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

$$\mathcal{L}(x) = f(x) + g(x), \quad \underset{x}{\operatorname{argmin}} \mathcal{L}(x) \neq \emptyset.$$

Then taking $\zeta_k = \frac{1}{\beta}$, x^k : k 'th PGD iterates.

$$\Rightarrow \mathcal{L}(x^k) - \min_x \mathcal{L}(x) \leq \frac{\beta \|x^1 - x_*\|_2^2}{2k}$$

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

APGD/FISTA (Accelerated proximal gradient descent,
Fast iterative thresholding algorithm).

Nesterov acceleration

[Beck and Teboulle, 2008]

$$\mu_1 = 0, \quad \mu_k = \frac{1 + \sqrt{1 + 4\mu_{k-1}^2}}{2}, \quad \text{and} \quad r_k = \frac{1 - \mu_k}{\mu_{k+1}}.$$

$$\begin{cases} v^{k+1} = \underset{x}{\operatorname{argmin}} \left\{ \frac{\beta}{2} \|x - (x^k - \frac{1}{\beta} \nabla f(x^k))\|_2^2 + g(x) \right\} \\ x^{k+1} = (1 - r_k) v^{k+1} + r_k v^k \end{cases} \quad y^1 = x^1$$

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

Thm: Under the same conditions,

Let x^k be the APGD iterates, then

$$\mathcal{L}(x^k) - \min_x \mathcal{L}(x) \leq \frac{2\beta \|x^1 - x_*\|_2^2}{k^2}.$$

Remark: Convergence analysis doesn't rely on how A and y are generated.

③ The approximate algorithm passing algorithm for LASSO.

ISTA w/ Step size 1.

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

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

$$\begin{cases} x^{k+1} = \eta(x^k + A^T z^k; \theta_k) \\ z^k = y - Ax^k + \underbrace{w_k z^{k-1}}_{\text{Onsager correction term. (Intuition will be given later)}} \end{cases} \quad \theta_k, w_k \text{ scalars.}$$

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

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

Remark: ⓧ Unlike ISTA, AMP is not always a descent algorithm. There is not a deterministic proof for convergence of AMP.

ⓧ Although the definition of AMP doesn't depend on any assumption on A and y , its theoretical analysis and performance highly depends on assumptions of A and y .

ⓧ Not as practical as ISTA / FISTA. AMP is more often used as a tool of theoretical analysis instead of a practical algorithm.

Assumptions:

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

b) $x_0 \in \mathbb{R}^d$, $\frac{1}{d} \sum_{i=1}^d \delta_{x_{0,i}} \xrightarrow{\text{weak}} P_{x_0}$, $\frac{1}{d} \sum_{i=1}^d x_{0,i}^2 \rightarrow \mathbb{E}_{x_0}[x_0^2] > 0$

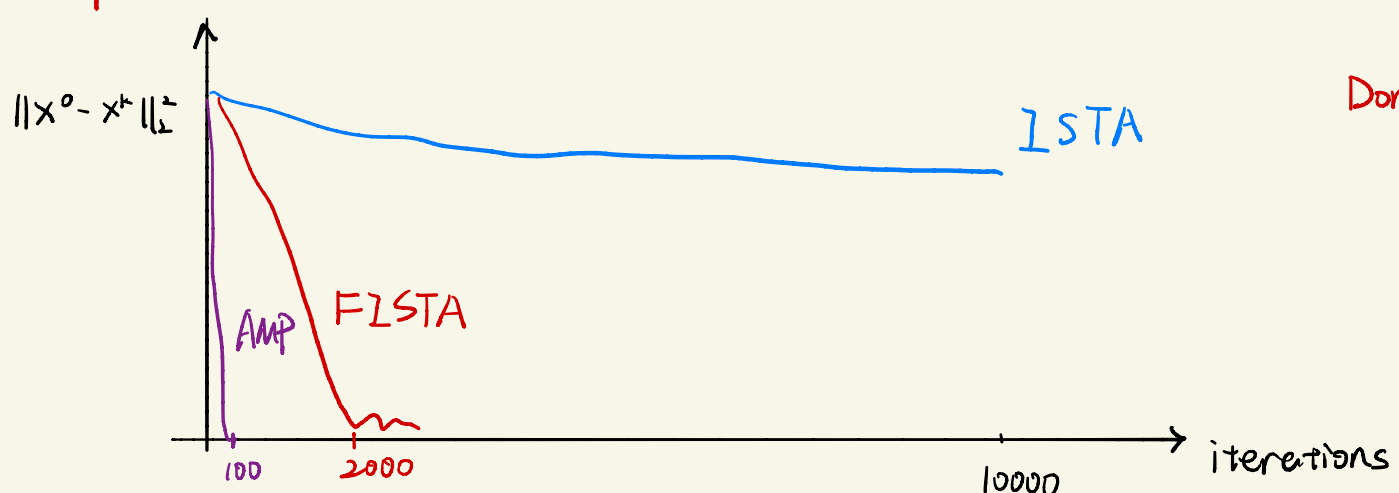
c) $w \in \mathbb{R}^n$, $\frac{1}{n} \sum_{i=1}^n \delta_{w_i} \xrightarrow{\text{weak}} P_w$, $\frac{1}{n} \sum_{i=1}^n w_i^2 \rightarrow \mathbb{E}_w[w^2] = \sigma^2$.

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

e) $n/d \rightarrow \delta$

Performance comparison.

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



Donoho's slides

④ Theoretical analysis of AMP.

State evolution characterization:

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

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

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

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

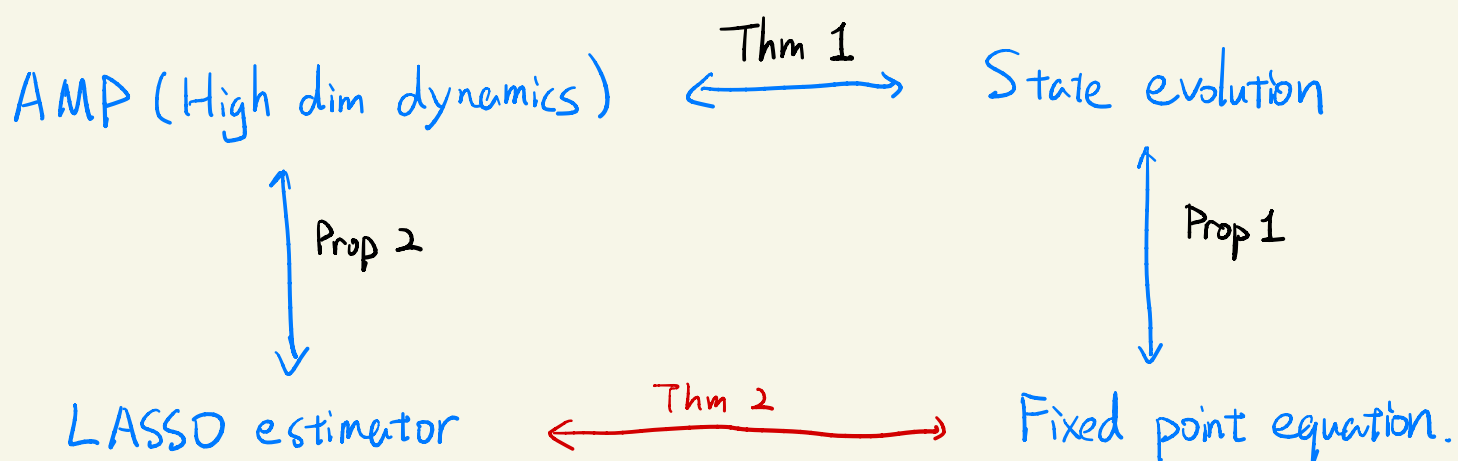
Let assumptions a) – e) holds, Let $\psi: \mathbb{R}^2 \rightarrow \mathbb{R}$ be any pseudo-Lipschitz function. i.e. $|\psi(x) - \psi(y)| \leq K \cdot \|x - y\|_2 \cdot (1 + \|x\|_2 + \|y\|_2)$.

Then almost surely

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

$$\text{Rmk: } \delta(\tau_{k+1}^2 - \sigma^2) = \lim_{d \rightarrow \infty} \|x^k - x_0\|_2^2 / d \quad \text{MSE at } k\text{-th iteration}$$

We will give the proof of this thm in later lectures.



Let (τ_*, α_*) be a proper solution of the Fixed point equation.
 there is a rigorous treatment

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

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

$$\tau_{k+1}^2 = F(\tau_k^2; \alpha_* \tau_k) \quad \text{converges to } \tau_*^2.$$

Moreover, convergence is exponentially fast.

Proof is based on analyzing properties of function F .

Prop 2: Let $\hat{x}(\lambda)$ be the LASSO estimator
and $\{x^k\}$ the AMP iterates. Then

$$\lim_{k \rightarrow \infty} \lim_{d \rightarrow \infty} \|\hat{x} - x^k\|_2^2 / d = 0.$$

Proof strategy:

⊗ Use convexity of $\mathcal{L}$ to show that

$$\exists v_k \in \partial \mathcal{L}(x^k), \quad \lim_{k \rightarrow \infty} \lim_{d \rightarrow \infty} \|v_k\|_2^2 / d = 0$$

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

$$\otimes \text{ Show } \exists v_k \in \partial \mathcal{L}(x^k), \quad \lim_{k \rightarrow \infty} \lim_{d \rightarrow \infty} \|v_k\|_2^2 / d = 0$$

using Thm 1. □

Concatinating Thm 1. Prop 1. Prop 2. we have

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

Let assumptions a) – e) holds, Let $\psi: \mathbb{R}^2 \rightarrow \mathbb{R}$ 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

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

Remark: Similar analysis and results work for a wide range
of problems, including $\mathbb{Z}_2$ synchronization, generalized Linear models.

Plans for remaining lectures:

⊗ Derivation of AMP algorithm.

⊗ Proof of AMP - SE relationship. (Thm 1)

⊗ Applications.