4th year Review - Statistical Physics Perspectives on Learning in High Dimensions

Advisor: Surya Ganguli

Madhu Advani

Stanford University

April 8, 2014

Outline

Current Research: High Dimensional M-estimation

1. Optimal Unregularized M-estimation
2. Optimal Regularized M-estimation

Future Directions

1. Extensions of High Dimensional M-estimation
2. Network Implementation of Learning

Introduction High Dimensional Inference

P = Number of Dimensions (Predictors)
N = Number of data points

$$\kappa = P/N$$

Classical Statistics

$P = O(1)$

$N \to \infty$

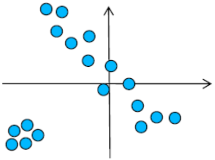
$P/N = \kappa \to 0$

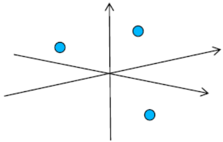
Modern Statistics

$P = O(N)$

$N \to \infty$

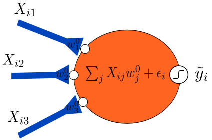
$P/N = \kappa \to O(1)$





Problem Setup

$$y_i = \mathbf{X_i} \cdot \mathbf{w^0} + \epsilon_i \qquad i \in [1, \dots N]$$



- Noise $\epsilon_i \sim f$ not necessarily gaussian
- (Easy) Classical Regime: $\kappa = P/N \to 0$
- (Hard) High Dimensional Regime: $\kappa = P/N \neq 0$

We want to find $\mathbf{w^0}$

Section 1

Unregularized Theory

M-estimation and Maximum Likelihood

M-estimation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[\sum_i \rho\left(y_i - \mathbf{X}_i \cdot \mathbf{w}\right) \right]$$

E.g. $\rho(x) = x^2, |x|, -\log f(x)$

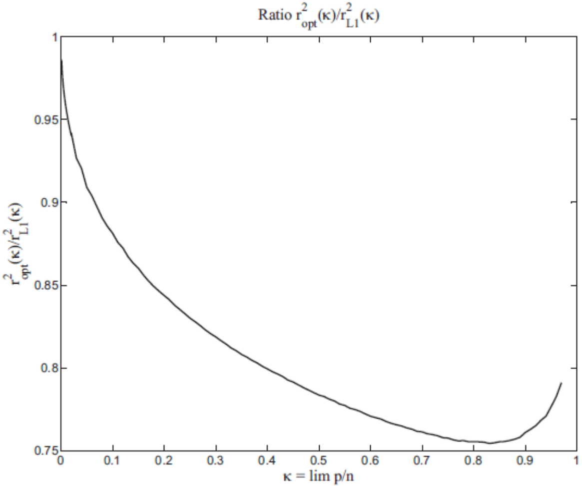
P fixed, $N \to \infty$

$$||\hat{w} - w^0||^2 = \frac{\left(\int \psi(x) f(x) dx\right)^2}{\int \psi^2 f(x) dx} \qquad \psi = \frac{\partial}{\partial w}\rho$$

$\psi_{\text{opt}} = \frac{f'}{f}$, $\rho_{\text{opt}} = -\log f$

$$\hat{\mathbf{w}}_{\text{ML}} = \arg\max_{\mathbf{w^0}} P\left(\mathbf{y}, \mathbf{X} | \mathbf{w^0}\right) = \arg\min_{\mathbf{w}} \left[\sum_i -\log f\left(y_i - \mathbf{X}_i \cdot \mathbf{w}\right) \right]$$

Sub-optimality of Maximum Likelihood in High Dimensions



El Karoui, et. al. PNAS 2013

Statistical Physics Formulation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_i \rho\left(y_i - \mathbf{X}_i \cdot \mathbf{w}\right)$$

Spin Glass System

Define spins $\mathbf{u} = \mathbf{w^0} - \mathbf{w}$ and Energy of system of spins:

$$E_{\mathbf{\Lambda}}(\mathbf{u}) = \sum_i \rho\left(\mathbf{X}_i \cdot \mathbf{u} + \epsilon_i\right)$$

Energy and temperature induce an equilibrium Gibbs distribution

$$P_G(\mathbf{u}) = \frac{e^{-\beta E_{\mathbf{\Lambda}}(\mathbf{u})}}{Z_{\mathbf{\Lambda}}} \qquad Z_{\mathbf{\Lambda}} = \int e^{-\beta E_{\mathbf{\Lambda}}(\mathbf{u})} du$$

$$\lim_{\beta\to\infty} P_G(\mathbf{u}) = \delta\left(\mathbf{u} - \mathbf{w^0} + \hat{\mathbf{w}}\right)$$

The Unregularized Case

$$q = \frac{1}{P}\sum_{j=1}^{P} \langle\, u_j \,\rangle_{\mathrm{G}}^2$$

Coupled Equations Relating Order Parameters q, c

$$\left\langle\!\left\langle \left(\mathrm{prox}_{c\rho}\left(\sqrt{q}z+\epsilon\right) - \sqrt{q}z - \epsilon\right)^2 \right\rangle\!\right\rangle_{z,\epsilon} = \kappa q$$

$$\left\langle\!\left\langle\, \mathrm{prox}'_{c\rho}(\sqrt{q}z+\epsilon) \,\right\rangle\!\right\rangle_{z,\epsilon} = 1-\kappa$$

- $X_{ij} \in \mathcal{N}(0, 1/P)$
- ρ convex
- $\epsilon_i \sim f$ iid

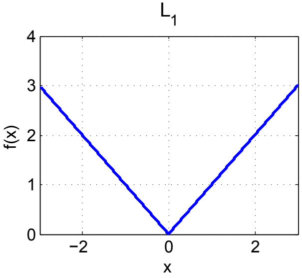
Proximal Map

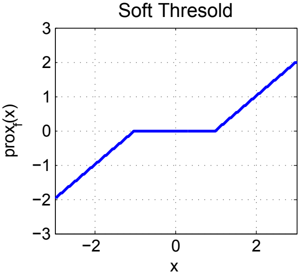
Definition

$$prox_f(x) = \arg\min_y \left[\frac{(x-y)^2}{2} + f(y)\right]$$

$$= [I + \partial f]^{-1}(x)$$

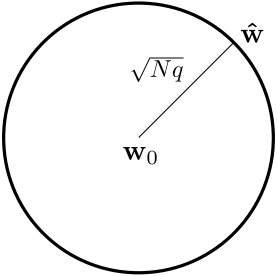
Example:





Analytic Estimator Error

How to choose ρ to minimize q?



$$\left\langle\left\langle \left(\mathrm{prox}_{c\rho}\left(\sqrt{q}z + \epsilon\right) - \sqrt{q}z - \epsilon\right)^2 \right\rangle\right\rangle_{z,\epsilon} = \kappa q$$

$$\left\langle\left\langle \mathrm{prox}'_{c\rho}(\sqrt{q}z + \epsilon) \right\rangle\right\rangle_{z,\epsilon} = 1 - \kappa$$

Optimal Unregularized M-estimator

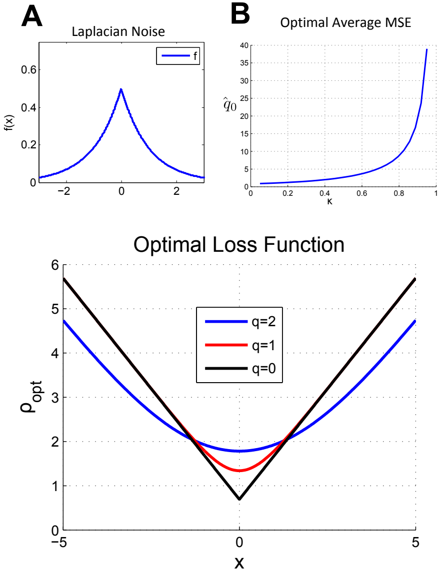
Optimal M-estimator

$$\rho_{\text{opt}}(x) = -\inf_y \left[\ln(\zeta_{\hat{q}}(y)) + \frac{(x-y)^2}{2\hat{q}} \right] \qquad \zeta_{\hat{q}} = f * \phi_{\hat{q}}$$

$$\hat{q} = \min q \quad \text{s.t.} \quad q I_q = \kappa \qquad I_q = \int_{-\infty}^{\infty} \frac{\zeta_q'^2(y)}{\zeta_q(y)} dy$$

- $\hat{q}$ - best possible asymptotic MSE for a convex M-estimator
- ρ_{opt} is the optimal loss function (assuming log-concave noise)
- Note: Not maximum likelihood, and ρ varies with dimensionality κ

Optimal Unregularized M-estimator



Section 2

Regularized Theory

Adding a Regularizer

$$P(\mathbf{w^0}|\mathbf{X},\mathbf{y}) = \frac{P(\mathbf{X},\mathbf{y}|\mathbf{w^0})P(\mathbf{w^0})}{P(\mathbf{X},\mathbf{y})} \propto \prod_i f(y_i - \mathbf{X_i}\cdot\mathbf{w^0}) \prod_j g(w_j^0)$$

Maximum a Priori

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min_{\mathbf{w}} \left[\sum_i -\log f(y_i - \mathbf{X}_i \cdot \mathbf{w}) + \sum_j -\log g(w_j) \right]$$

Regularized M-estimation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[\sum_i \rho(y_i - \mathbf{X}_i \cdot \mathbf{w}) + \sum_j \sigma(w_j) \right]$$

Note separability. Solvable for convex σ, ρ

Motivation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_a \rho(y_a - \mathbf{X_a} \cdot \mathbf{w}) + \sum_i \sigma(w_i)$$

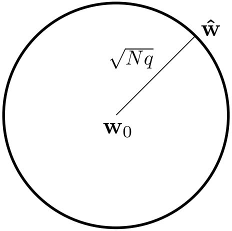
Applications

- Maximum Likelihood (ML) and MAP commonly applied to High Dimensional Bio-informatics problems, where we should expect poor performance
- Deriving/Understanding the Optimal M-estimator has potential applications for statistical inference and signal processing.
- Tractability of this form of optimization popular: Compressed Sensing, LASSO, Elastic Net

Regularized M-estimation

$$y_i = \mathbf{X}_i \cdot \mathbf{w^0} + \epsilon_i \qquad\qquad w_j^0 \sim g, \epsilon_i \sim f$$

$$E_{\mathbf{\Lambda}}(\mathbf{u}) = \sum_i \rho\left(\mathbf{X}_i \cdot \mathbf{u} + \epsilon_i\right) + \sum_a \sigma\left(w_a^0 - u_a\right)$$



Regularized M-estimation

Optimal Inference

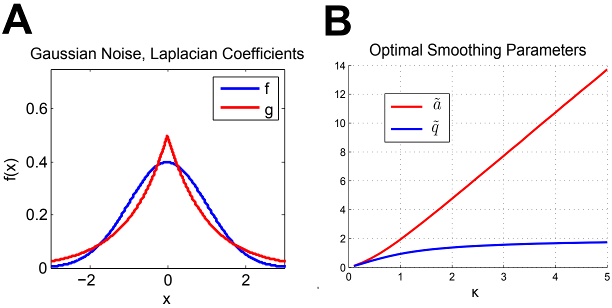
$$\rho_{\mathrm{opt}}^{R}(x) = -\inf_{y}\left[\ln(\zeta_{\tilde{q}_0}(y)) + \frac{(x-y)^2}{2\tilde{q}_0}\right]$$

$$\sigma_{\mathrm{opt}}^{R}(x) = -\frac{\tilde{q}_0}{\tilde{a}}\inf_{y}\left[\ln(\xi_{\tilde{a}}(y)) + \frac{(x-y)^2}{2\tilde{a}}\right]$$

$$\tilde{q}_0, \tilde{a} = \arg\min_{q_0, a} q_0$$

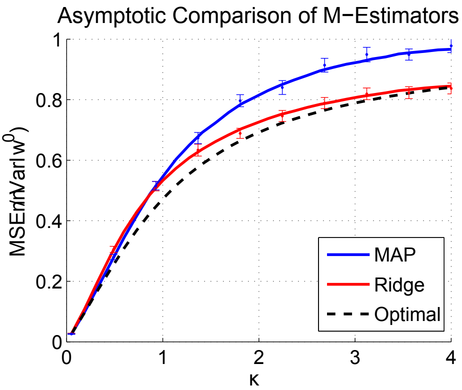
$$\text{s.t.} \quad aI_{q_0} = \kappa, \quad a^2 J_a = a - q_0 \qquad I_q = \int \frac{\zeta_q'^2}{\zeta_q}, \quad J_a = \int \frac{\xi_a'^2}{\xi_a}$$

- $\rho_{\mathrm{opt}}, \sigma_{\mathrm{opt}}$ are optimal M-estimator (log concave f, g)
- $\tilde{q}_0$ is the asymptotic MSE
- $\tilde{q}_0, \tilde{a}$ are smoothing parameters


A
Gaussian Noise, Laplacian Coefficients
f
g
0.6
0.4
0.2
0
f(x)
−2
0
2
x
B
Optimal Smoothing Parameters
14
12
10
8
6
4
2
0
$\tilde{a}$
$\tilde{q}$
0
1
2
3
4
5
κ

M-estimator Comparison

$$\rho_{\text{MAP}}(x) = \frac{x^2}{2} \qquad \sigma_{\text{MAP}}(x) = |x|$$
$$\rho_{\text{Ridge}}(x) = \frac{x^2}{2} \qquad \sigma_{\text{Ridge}}(x) = \frac{x^2}{2}$$



Section 3

Future Extensions of High Dimensional M-estimation

Generalized Models

$$y_i = \eta(\mathbf{X_i} \cdot \mathbf{w^0}) + \epsilon_i$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[\sum_i \rho(y_i - \eta\left(\mathbf{X_i} \cdot \mathbf{w}\right)) + \sum_j \sigma(w_j) \right]$$

Energy Function

$$E(\mathbf{w}) = \sum_i \rho\left(\eta\left(\mathbf{X_i} \cdot \mathbf{w^0}\right) - \eta\left(\mathbf{X_i} \cdot (\mathbf{w^0} - \mathbf{u}) + \epsilon_i\right)\right) + \sum_j \sigma(w_j^0 - u_j)$$

A Theory for Non log-concave Noise?

- For non log-concave noise, we cannot naively expect ρ_{opt} to be optimal, because it may be non-convex and thus violate the RS assumptions used to derive it.

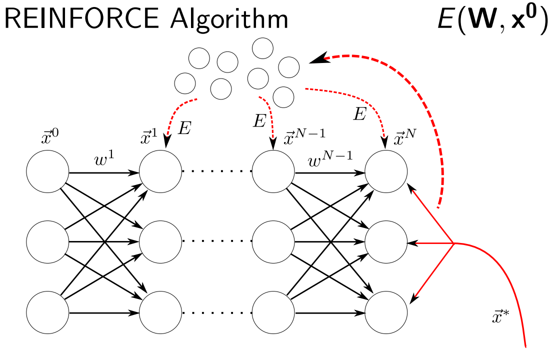
- An alternative perform a numerical optimization over a parameterized set of convex functions.

- This same concept could be applied to find better alternatives to compressed sensing, or other convex inference algorithms

Section 4

Possible Network Learning Algorithms

Local Reinforcement Learning through Noise Injection

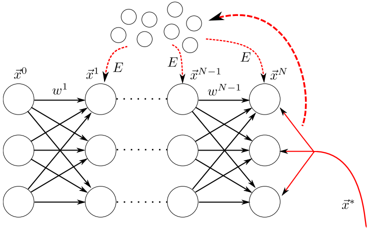


Injecting Noise η into a single synapse W_i gives

$$\Delta E \approx \frac{\partial E}{\partial W_i}\eta \implies \langle \eta \Delta E \rangle = \frac{\partial E}{\partial W_i}$$

Local Reinforcement Learning through Noise Injection

REINFORCE Algorithm $E(\mathbf{W}, \mathbf{x^0})$



Number of Synapses = S

- Adding noise to each synapse individually O(S) updates
- Adding noise to every synapse increases SNR. Fails catastrophically requiring integrating signals over O(S) updates
- One idea is to use reinforcement learning to intelligently modify weights based on past Errors and Actions

Acknowledgements

Committee

Surya Ganguli
Daniel Fisher
Stephen Baccus

Ganguli Lab

Subhaneil Lahiri
Jascha Sohl-Dickstein
Peiran Gao
Niru Maheswaranathan
Ben Poole
Kiah Hardcastle

Funding

Stanford Graduate Fellowship
Mind Brain Computation IGERT