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

Advisor: Surya Ganguli

Madhu Advani

Stanford University

April 7, 2014

Outline

Current Research

Optimal Tractable High Dimensional M-estimation

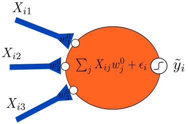
Future Directions

1. LN and GLM extensions of M-estimation also Optimal Signal Processing (Structured Coefficients)
2. Random Dimensionality Reduction
3. Phase transitions in clustering Behavior

Previous Research and Future Plan

1. Review Paper
2. Timeline

Problem Setup



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

- 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 $\mathbf{w^0}$

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

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

Maximum Likelihood

P fixed, $N \to \infty$

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

$\psi = \frac{\partial}{\partial w} \rho$

We find that $\rho_{\text{opt}} = -\log f$

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

Classical vs High Dimensional Optimal M-estimation

Classical

- $N \to \infty$, $P/N = \kappa \to 0$
- $\rho_{\text{opt}} = -\log f$
- $\langle\langle\, (\hat{w}_i - w_i^0)^2 \,\rangle\rangle \geq \frac{\kappa}{\int \frac{f'^2}{f}}$

High Dimensional

- $N, P \to \infty$, $P/N = \kappa \in [0, 1]$
- $\rho_{\text{opt}}(x) = -\inf_y \left[\ln(\zeta(y)) + \frac{(x-y)^2}{2\hat{q}_0} \right]$ $\qquad \zeta = f * \phi_{\hat{q}_0}$
- $\langle\langle\, (\hat{w}_i - w_i^0)^2 \,\rangle\rangle \geq \frac{\kappa}{\int \frac{\zeta'^2}{\zeta}}$

Adding a Regularizer

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

$w_j^0 \sim g$

Maximum a Priori

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

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

Statistical Physics Formulation

Define a spin glass system to solve M-estimator inference

Spin Glass System

Define continuous spins $\mathbf{u} = \mathbf{w^0} - \mathbf{w}$. Let the Energy of the system be a function of these spins

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

This in turn induces an equilibrium probability distribution on the state

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

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

The Unregularized Case

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

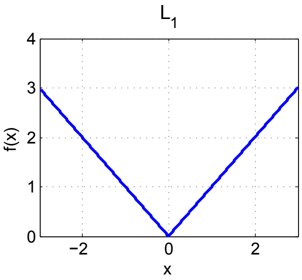
Proximal Map

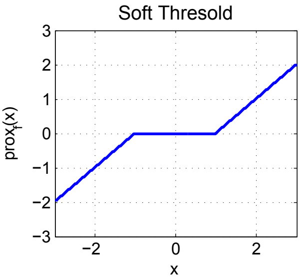
Definition

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

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

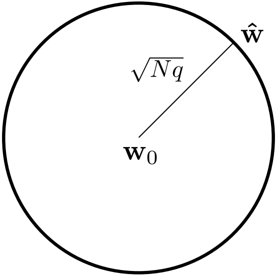
Example:





Analytic Estimator Error

How to choose ρ to minimize q?



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

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

Optimal Unregularized M-estimator

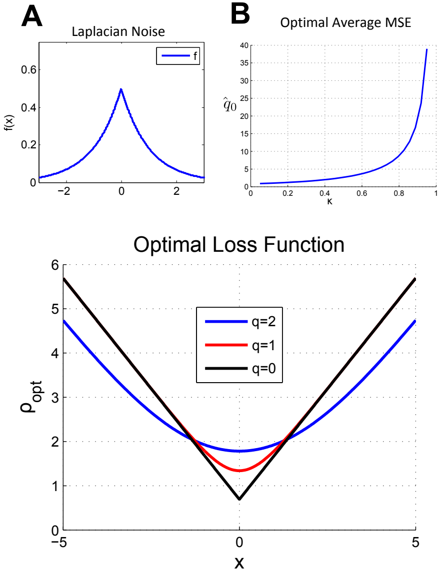
Optimal M-estimator

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

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

- $\hat{q}$ - 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

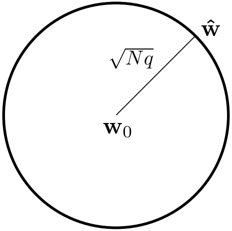
Optimal Unregularized M-estimator



Regularized M-estimation

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

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



Regularized M-estimation

Optimal Inference

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

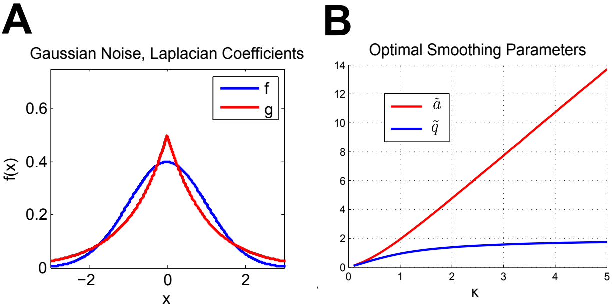
$$\sigma_{\text{opt}}^{R}(x) = -\frac{\tilde{q}_0}{\tilde{a}}\inf_{y}\left[\ln(\xi_{\tilde{a}}(y)) + \frac{(x-y)^2}{2\tilde{a}}\right]$$

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

$$\text{s.t.} \qquad a I_{q_0} = \kappa, \quad a^2 J_a = a - q_0$$

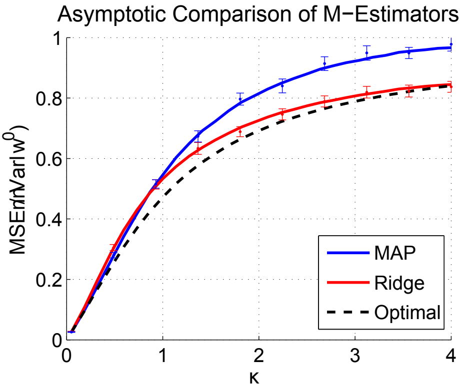
- $\rho_{\text{opt}}, \sigma_{\text{opt}}$ are optimal M-estimator (log concave f, g)
- $\tilde{q}_0$ is the asymptotic MSE
- $\tilde{q}_0, \tilde{a}$ are smoothing parameters

Unregularized M-estimator



M-estimator Comparison

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



Future Directions- LN models

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

Future Directions - Random Dimensionality Reduction

Future Directions - Phase Transitions in Clustering

Future Directions - Phase Transitions in Clustering