OOPS 2020
Mean field methods in high-dimensional statistics and nonconvex optimization
Lecturer: Andrea Montanari
Problem session leader: Michael Celentano
July 7, 2020

Problem Session 1

Problem 1: from Gordon's objective to the fixed point equations

Recall *Gordon's min-max problem* is

$$B^*(\boldsymbol{g},\boldsymbol{h}) := \min_{\boldsymbol{u}\in\mathbb{R}^d}\max_{\boldsymbol{v}\in\mathbb{R}^n}\left\{\frac{1}{n}\|\boldsymbol{v}\|\langle\boldsymbol{g},\boldsymbol{u}\rangle + \frac{1}{n}\|\boldsymbol{u}\|\langle\boldsymbol{h},\boldsymbol{v}\rangle - \frac{\sigma}{n}\langle\boldsymbol{w},\boldsymbol{v}\rangle - \frac{1}{2n}\|\boldsymbol{v}\|^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{\theta}_0+\boldsymbol{u}\|_1\right\}. \tag{1}$$

In lecture, we claimed that by analyzing Gordon's objective we can show that the Lasso solution is described in terms of the solutions τ^*,β^* to the *fixed point equations*

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left[(\eta(\Theta+\tau Z;\tau\lambda/\beta)-\Theta)^2\right],$$
$$\beta = \tau\left(1-\frac{1}{\delta}\mathbb{E}[\eta'(\Theta+\tau Z;\tau\lambda/\beta)\right),$$

where η is the solution of the 1-dimensional problem

$$\eta(y;\alpha) := \arg\min_{x\in\mathbb{R}}\left\{\frac{1}{2}(y-x)^2+\alpha|x|\right\} = (|x|-\alpha)_+\mathsf{sign}(x).$$

η is commonly known as *soft-thresholding*. In this problem, we will outline how to derive the fixed point equations from Gordon's min-max problem.

(a) Define $\tilde{\boldsymbol{\theta}}_0 = \sqrt{n}\boldsymbol{\theta}_0$ and $\tilde{\boldsymbol{u}} = \sqrt{n}\boldsymbol{u}$. Prove that $B^*(\boldsymbol{g}, \boldsymbol{h})$ has the same distribution as

$$\min_{\tilde{\boldsymbol{u}}\in\mathbb{R}^d} \max_{\beta\geq 0} \left\{ \left(\left\| \frac{\|\tilde{\boldsymbol{u}}\|}{\sqrt{n}} \frac{\boldsymbol{h}}{\sqrt{n}} - \frac{\sigma \boldsymbol{w}}{\sqrt{n}} \right\| - \frac{1}{n}\langle \boldsymbol{g}, \tilde{\boldsymbol{u}} \rangle \right) \beta - \frac{1}{2}\beta^2 + \frac{\lambda}{n}\|\tilde{\boldsymbol{\theta}}_0 + \tilde{\boldsymbol{u}}\|_1 \right\}. \qquad (2)$$

Hint: Let $\beta = \|\boldsymbol{v}\|/\sqrt{n}$

$$\frac{1}{n^{3/2}}\|v\|\langle g, \tilde{u}\rangle + \underbrace{\frac{1}{n^{3/2}}\|\tilde{u}\|\langle h, v\rangle - \frac{\sigma}{n}\langle w, v\rangle} - \frac{1}{2n}\|v\|^2 + \frac{\lambda}{n}\|\tilde{\theta}_0 + \tilde{u}\|_1$$

$$\left\langle \frac{\|\tilde{u}\|}{n^{1/2}} \frac{h}{n^{1/2}} - \frac{\sigma w}{n^{1/2}}), \frac{v}{n^{1/2}} \right\rangle$$

$$\underbrace{\left\| \frac{\|\tilde{u}\|}{n^{1/2}} \frac{h}{n^{1/2}} - \frac{\sigma w}{n^{1/2}} \right\|}_{\sqrt{\frac{\|\tilde{u}\|^2}{n} + \sigma^2}} \underbrace{\frac{\|v\|}{n^{1/2}}}_{\beta}$$

(b) Argue (heuristically) that we may approximate the optimization above by

$$\min_{\tilde{\boldsymbol{u}}\in\mathbb{R}^d}\max_{\beta\geq 0}\left\{\left(\sqrt{\frac{\|\tilde{\boldsymbol{u}}\|^2}{n}+\sigma^2}-\frac{1}{n}\langle \boldsymbol{g},\boldsymbol{u}\rangle\right)\beta-\frac{1}{2}\beta^2+\frac{\lambda}{n}\|\tilde{\boldsymbol{\theta}}_0+\tilde{\boldsymbol{u}}\|_1\right\}. \qquad (3)$$

Remark: If we maximize over $\beta \geq 0$ explicitly, we get

$$\min_{\tilde{\boldsymbol{u}}\in\tilde{U}}\left\{\frac{1}{2}\left(\sqrt{\frac{\|\tilde{\boldsymbol{u}}\|^2}{n}+\sigma^2}-\frac{\langle \boldsymbol{g},\tilde{\boldsymbol{u}}\rangle}{n}\right)_+^2+\frac{\lambda}{n}\|\tilde{\boldsymbol{\theta}}_0+\tilde{\boldsymbol{u}}\|_1\right\}.$$

Note that the objective on the right-hand side is locally strongly convex around any point $\tilde{\boldsymbol{u}}$ at which the first term is positive. When $n < d$, the Lasso objective is nowhere locally strongly convex. This convenient feature of the new form of Gordon's problem is very useful for its analysis. We do not explore this further here.

$$\frac{1}{2}\|y-X\theta\|^2+\lambda\|\theta\|_1$$

$$n < d$$

$$\sqrt{x} = \min_{\tau\geq 0}\ \frac{x}{2\tau}+\frac{\tau}{2}$$

$$\max_{\beta\geq 0}\ \min_{\tilde{u}\in\mathbb{R}^d}\Big\{\Big(\min_{\tau\geq 0}\ \frac{\|\tilde{u}\|/n+\sigma^2}{2\tau}+\frac{\tau}{2}\Big)-\frac{1}{n}\langle g,\tilde{u}\rangle\Big)\beta$$

$$-\frac{1}{2}\beta^2+\frac{\lambda}{n}\|\tilde{\theta}_0+\tilde{u}\|_1\Big\}$$

$$\max_{\beta\geq 0}\ \min_{\tau\geq 0}\Big\{\frac{\sigma^2\beta}{2\tau}+\frac{\tau\beta}{2}-\frac{1}{2}\beta^2+\frac{1}{n}\min_{\tilde{u}\in\mathbb{R}^d}\Big\{\frac{\beta}{2\tau}\|\tilde{u}\|^2-\beta\langle g,\tilde{u}\rangle$$

$$+\frac{\lambda}{n}\|\tilde{\theta}_0+\tilde{u}\|_1\Big\}\Big\}$$

(c) Argue that the quantity in Eq. (3) is equal to

$$\max_{\beta\geq 0}\min_{\tau\geq 0}\left\{\frac{\sigma^2\beta}{2\tau}+\frac{\tau\beta}{2}-\frac{\beta^2}{2}+\frac{1}{n}\min_{\tilde{\boldsymbol{u}}\in\mathbb{R}^d}\left\{\frac{\beta}{2\tau}\|\tilde{\boldsymbol{u}}\|^2-\beta\langle\boldsymbol{g},\tilde{\boldsymbol{u}}\rangle+\lambda\|\tilde{\boldsymbol{\theta}}_0+\tilde{\boldsymbol{u}}\|_1\right\}\right\}. \qquad (4)$$

Hint: Recall the identity $\sqrt{x}=\min_{\tau>0}\left\{\frac{x}{2\tau}+\frac{\tau}{2}\right\}$. $x \geq 0$

$\tau \geq 0$

$$\underbrace{\frac{1}{d}\sum_{j=1}^{d}\left\{\frac{\beta}{2\tau}\hat{v}_j^2-\beta g_j v_j+\lambda|\hat{\theta}_{0j}+\hat{v}_j|\right\}}$$

$$\frac{1}{d}\sum_{j=1}^{d}\delta_{(\hat{v}_j,\,g_j,\,\hat{\theta}_{0j})}\xrightarrow{w}\mu$$

$\underline{\text{wrt } \tau}$ $$-\frac{\sigma^2\cancel{\beta}}{\cancel{2}\tau^2}+\frac{\cancel{\beta}}{\cancel{2}}-\underbrace{\frac{\cancel{\beta}}{2n\tau^2}\mathbb{E}\left[\|\hat{v}(g)\|^2\right]}_{-\frac{1}{\cancel{2}\tau^2}\cdot\frac{1}{\delta}\mathbb{E}\left[(\eta(\Theta+\tau Z;\lambda\tau/\beta)-\Theta)^2\right]}=0$$ soln to part (e).

$$\tau^2=\sigma^2+\frac{1}{\delta}\underbrace{\mathbb{E}\left[|\eta(\Theta+\tau Z;\lambda\tau/\beta)-\Theta)^2\right]}_{}$$

$\underline{\text{wrt } \beta}$

$$\frac{\sigma^3}{2\tau}+\frac{\tau}{2}-\beta+\underbrace{\frac{1}{2n\tau}\mathbb{E}\left[\|\hat{v}\|^2\right]}_{\frac{1}{2}\frac{1}{\tau}(\tau^2-\sigma^2)}-\frac{1}{n}\mathbb{E}\left[\langle g,\hat{v}(g)\rangle\right]$$

$$-\frac{\tau}{\delta}\mathbb{P}\left(|\Theta+\tau Z|\geq\lambda\tau/\beta\right)=0$$

$$\beta=\tau\left(1-\frac{1}{\delta}\mathbb{P}\left(|\Theta+\tau Z|\geq\lambda\tau/\beta\right)\right)$$

(d) Write

$$\widehat{\boldsymbol{u}} := \arg\min_{\tilde{\boldsymbol{u}}\in\mathbb{R}^d} \left\{ \frac{\beta}{2\tau}\|\tilde{\boldsymbol{u}}\|^2 - \beta\langle \boldsymbol{g}, \tilde{\boldsymbol{u}}\rangle + \lambda\|\tilde{\boldsymbol{\theta}}_0 + \tilde{\boldsymbol{u}}\|_1 \right\}$$

in terms of the soft-thresholding operator.

(e) Compute

(i) the derivative of $\min_{\tilde{\boldsymbol{u}}\in\mathbb{R}^d} \left\{ \frac{\beta}{2\tau}\|\tilde{\boldsymbol{u}}\|^2 - \beta\langle \boldsymbol{g}, \tilde{\boldsymbol{u}}\rangle + \lambda\|\tilde{\boldsymbol{\theta}}_0 + \tilde{\boldsymbol{u}}\|_1 \right\}$ with respect to τ.

(ii) the derivative of $\min_{\tilde{\boldsymbol{u}}\in\mathbb{R}^d} \left\{ \frac{\beta}{2\tau}\|\tilde{\boldsymbol{u}}\|^2 - \beta\langle \boldsymbol{g}, \tilde{\boldsymbol{u}}\rangle + \lambda\|\tilde{\boldsymbol{\theta}}_0 + \tilde{\boldsymbol{u}}\|_1 \right\}$ with respect to β.

$$\frac{\beta}{\tau}\left(\frac{1}{2}\|\tilde{u} - \tau g\|^2 - \cancel{\frac{\tau^2\|g\|^2}{2}} + \frac{\lambda\tau}{\beta}\|\tilde{\theta}_0 + \tilde{u}\|_1 \right)$$

$$\theta = \tilde{\theta}_0 + \tilde{u}$$

$$\frac{1}{2}\|\theta - (\tilde{\theta}_0 + \tau g)\|^2 + \frac{\lambda\tau}{\beta}\|\theta\|_1$$

$$\hat{u} = \eta(\tilde{\theta}_0 + \tau g; \lambda\tau/\beta) - \tilde{\theta}_0$$

(e) $\frac{d}{d\beta}\min_x C(x;\beta) = \partial_\beta C(\hat{x};\beta)$

$$C(\hat{x}(\beta);\beta)$$

$$\cancelto{0}{\nabla_x C(\hat{x};\beta)^T dx} + \partial_\beta C(\hat{x}(\beta);\beta)\, d\beta$$

Derivative wrt τ: $\quad -\frac{\beta}{2\tau^2}\|\hat{u}(g)\|^2$

Derivative wrt β: $\quad \frac{1}{2\tau}\|\hat{u}(g)\|^2 - \langle g, \hat{u}(g)\rangle$

(f) Write $\frac{1}{n}\mathbb{E}[\|\widehat{\boldsymbol{u}}\|^2]$ and $\frac{1}{n}\mathbb{E}[\langle \boldsymbol{g}, \widehat{\boldsymbol{u}}\rangle]$ as an expectation over the random variables $(\Theta, Z) \sim \widehat{\mu}_{\boldsymbol{\theta}_0} \otimes \mathsf{N}(0,1)$. For the latter, rewrite it using Gaussian integration by parts.

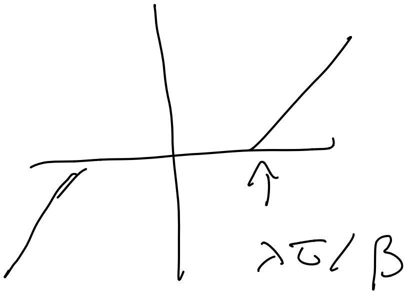
(g) Take the derivative of the objective in Eq. (4) with respect to τ and with respect to β. Show that setting the expectations of these derivatives to 0 is equivalent to the fixed point equations.

$$\frac{1}{n}\mathbb{E}[\|\hat{u}(g)\|^2]$$

$$= \frac{d}{n}\cdot\frac{1}{d}\mathbb{E}_g\Big[\underbrace{\|\eta(\tilde{\theta}_0+\tau g;\lambda\tau/\beta)-\tilde{\theta}_0\|^2}_{\sum_{j=1}^{d}\left(\eta(\hat{\theta}_{0j}+\tau g_j;\lambda\tau/g)-\tilde{\theta}_{0j}\right)^2}\Big]$$

$$\approx \frac{d}{n}\mathbb{E}_{\hat{\mu}}\left[\left(\eta(\Theta+\tau Z;\lambda\tau/\beta)-\Theta\right)^2\right]$$

$$\hat{\mu} = \frac{1}{d}\sum_{j=1}^{d}\delta_{(\tilde{\theta}_{0j},\, g_j)}$$



$$\xrightarrow{w} \mu_{\Theta}\otimes N(0,1)$$

$$\frac{1}{n}\mathbb{E}[\langle g, \eta(\hat{\theta}_0+\tau g;\lambda\tau/\beta)\rangle]$$

$$\frac{d}{n}\mathbb{E}_{\hat{\mu}}[Z\,\eta(\Theta+\tau Z;\lambda\tau/\beta)]$$

$$\frac{d}{n}\tau\mathbb{E}\left[\mathbb{1}\left\{|\textcircled{n}+\tau z| \geq \lambda\tau/\beta\right\}\right]$$

$$= \frac{d}{n}\tau\mathbb{P}\left(|\textcircled{n}+\tau z| \geq \frac{\lambda\tau}{\beta}\right)$$

Problem 2: Gordon's objective for max-margin classification

The use of Gordon's technique extends well beyond linear models. In logistic regression, we receive iid samples according to

$$y_i \sim \mathsf{Rad}(f(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_0\rangle)), \quad \boldsymbol{x}_i \sim \mathsf{N}(0, \mathbf{I}_d),$$
$$f(x) = \frac{\exp(x)}{\exp(x) + \exp(-x)}.$$

For a certain $\delta^* > 0$ the following occurs: when $n/p \to \delta < \delta^*$, $n, p \to \infty$, with high probability there exists $\boldsymbol{\theta}$ such that $y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle > 0$ for all $i = 1, \dots, n$. In such a regime, the data is *linearly separable.* The max-margin classifier is defined as

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}} \left\{ \min_{i \leq n} y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle : \|\boldsymbol{\theta}\| \leq 1 \right\}, \tag{5}$$

and the value of the optimization problem, which we denote by $\kappa(\boldsymbol{y}, \boldsymbol{X})$, is called the *maximum margin.* To simplify notation, we assume in this problem that $\|\boldsymbol{\theta}_0\| = 1$. In this problem, we outline how to set up Gordon's problem for max-margin classification. The analysis of Gordon's objective is complicated, and we do not describe it here. See Montanari, Ruan, Sohn, Yan (2019+). "The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparameterized regime." `arxiv:1911.01544`.

(a) Show that

$$\kappa(\boldsymbol{y}, \boldsymbol{X}) \geq \kappa \text{ if and only if } \min_{\|\boldsymbol{\theta}\| \leq 1} \|(\kappa \mathbf{1} - (\boldsymbol{y} \odot \boldsymbol{X}\boldsymbol{\theta}))_+\|_2 = 0.$$

Argue that

$$\begin{aligned} \min_{\|\boldsymbol{\theta}\| \leq 1} \frac{1}{\sqrt{d}} \|(\kappa \mathbf{1} - (\boldsymbol{y} \odot \boldsymbol{X}\boldsymbol{\theta}))_+\|_2 &= \min_{\|\boldsymbol{\theta}\| \leq 1} \max_{\|\boldsymbol{\lambda}\| \leq 1, \boldsymbol{\lambda} \geq 0} \frac{1}{\sqrt{d}} \boldsymbol{\lambda}^\top (\kappa \mathbf{1} - \boldsymbol{y} \odot \boldsymbol{X}\boldsymbol{\theta}) \\ &= \min_{\|\boldsymbol{\theta}\| \leq 1} \max_{\|\boldsymbol{\lambda}\| \leq 1, \boldsymbol{\lambda} \odot \boldsymbol{y} \geq 0} \frac{1}{\sqrt{d}} \boldsymbol{\lambda}^\top (\kappa \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}). \end{aligned}$$

(b) Why can't we use Gordon's inequality to compare the preceding min-max problem to

$$\min_{\|\boldsymbol{\theta}\| \leq 1} \max_{\|\boldsymbol{\lambda}\| \leq 1, \boldsymbol{y} \odot \boldsymbol{\lambda} \geq 0} \frac{1}{\sqrt{d}} (\kappa \boldsymbol{\lambda}^\top \boldsymbol{y} + \|\boldsymbol{\lambda}\| \boldsymbol{g}^\top \boldsymbol{\theta} + \|\boldsymbol{\theta}\| \boldsymbol{h}^\top \boldsymbol{\lambda})?$$

(c) Let $\tilde{\boldsymbol{x}} = \boldsymbol{X}\boldsymbol{\theta}_0$. Show that the min-max problem is equivalent to

$$\min_{\|\boldsymbol{\theta}\|\le 1} \max_{\|\boldsymbol{\lambda}\|\le 1, \boldsymbol{\lambda}\ge 0} \frac{1}{\sqrt{d}} \boldsymbol{\lambda}^\top (\kappa \mathbf{1} - (\boldsymbol{y} \odot \tilde{\boldsymbol{x}}) \langle \boldsymbol{\theta}_0, \boldsymbol{\theta} \rangle - \boldsymbol{y} \odot \boldsymbol{X} \Pi_{\boldsymbol{\theta}_0^\perp} \boldsymbol{\theta}).$$

Here $\Pi_{\boldsymbol{\theta}_0^\perp}$ is the projection operator onto the orthogonal complement of the space spaned by $\boldsymbol{\theta}_0$. Argue that

$$\begin{aligned} &\mathbb{P}\left(\min_{\|\boldsymbol{\theta}\|\le 1} \max_{\|\boldsymbol{\lambda}\|\le 1, \boldsymbol{\lambda}\ge 0} \frac{1}{\sqrt{d}} \boldsymbol{\lambda}^\top (\kappa \mathbf{1} - (\boldsymbol{y} \odot \tilde{\boldsymbol{x}}) \langle \boldsymbol{\theta}_0, \boldsymbol{\theta} \rangle - \boldsymbol{y} \odot \boldsymbol{X} \Pi_{\boldsymbol{\theta}_0^\perp} \boldsymbol{\theta}) \le t \right) \\ &\qquad\le 2\mathbb{P}\left(\min_{\|\boldsymbol{\theta}\|\le 1} \max_{\|\boldsymbol{\lambda}\|\le 1, \boldsymbol{\lambda}\ge 0} \frac{1}{\sqrt{d}} \left(\boldsymbol{\lambda}^\top (\kappa \mathbf{1} - (\boldsymbol{y} \odot \tilde{\boldsymbol{x}}) \langle \boldsymbol{\theta}_0, \boldsymbol{\theta} \rangle) + \|\boldsymbol{\lambda}\| \boldsymbol{g}^\top \Pi_{\boldsymbol{\theta}_0^\perp} \boldsymbol{\theta} + \|\Pi_{\boldsymbol{\theta}_0^\perp} \boldsymbol{\theta}\| \boldsymbol{h}^\top \boldsymbol{\lambda} \right) \le t \right), \end{aligned}$$

and likewise for the comparison of the probabilities that the min-max values exceed t, where $\boldsymbol{g} \sim \mathsf{N}(0, \mathbf{I}_d)$ and $\boldsymbol{h} \sim \mathsf{N}(0, \mathbf{I}_n)$ independent of everything else.

(d) What is the limit in Wasserstein-2 distance of $\frac{1}{n}\sum_{i=1}^n \delta_{(y_i, \tilde{x}_i, h_i)}$?