These notes have not received the scrutiny of publication. They could be missing important references, etc.

Gordon Theorem and its Applications

1 Gaussian min - max

Theorem 1 (Gaussian min - max)**.** *Let X, Y be sets, and let $A \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d. entries $A_{ij} \sim \mathcal{N}(0,1)$. Let $g \sim \mathcal{N}(0, I_m)$, $h \sim \mathcal{N}(0, I_n)$, and $\tilde{g} \sim \mathcal{N}(0,1)$ be mutually independent Gaussian random variables.*

$$\Pr\left(\min_{x\in X}\max_{y\in Y}\left\{\,\langle y, Ax\rangle + \tilde{g}\,|x|\cdot|y| + \psi(x,y)\right\} \geq c\right) \geq \Pr\left(\min_{x\in X}\max_{y\in Y}\left\{\,(|x|\,\langle g,y\rangle + |y|\,\langle h,x\rangle + \psi(x,y))\,\right\} \geq c\right). \quad (1)$$

Remark 1. We refer to the left-hand side (LHS) of inequality (9) as the *Primary Optimization (PO)* problem, and to the right-hand side (RHS) as the *Auxiliary Optimization (AO)* problem.

Definition 1. Define

$$B_{x,y} = \langle y, Ax\rangle + \tilde{g}\;|x|\cdot|y|\,, \qquad D_{x,y} = |x|\,\langle g,y\rangle + |y|\,\langle h,x\rangle. \quad (2)$$

From above definition, we immediately have

$$\mathbb{E}[B_{x,y}^2] = |x|^2\,|y|^2 + |x|^2\,|y|^2 = \mathbb{E}[D_{x,y}^2]. \quad (3)$$

$$\mathbb{E}[D_{x,y}D_{x',y'}] = \mathbb{E}[(|x|\,\langle g,y\rangle + |y|\,\langle h,x\rangle)(|x'|\,\langle g,y'\rangle + |y'|\,\langle h,x'\rangle)] = |x|\,|x'|\,\langle y,y'\rangle + |y|\,|y'|\,\langle x,x'\rangle. \quad (4)$$

$$\mathbb{E}[B_{x,y}B_{x',y'}] = \mathbb{E}[(\langle y,Ax\rangle + \tilde{g}\,|x|\,|y|)(\langle y',Ax'\rangle + \tilde{g}\,|x'|\,|y'|)] = \langle x,x'\rangle\langle y,y'\rangle + |x|\,|x'|\,|y|\,|y'|\,. \quad (5)$$

Difference of correlations. By subtracting equation (4) and (5), we immediately have the following properties:

$$\mathbb{E}[D_{x,y}D_{x',y'} - B_{x,y}B_{x',y'}] = -(|x|\,|x'| - \langle x,x'\rangle)(|y|\,|y'| - \langle y,y'\rangle), \quad (6)$$

which is always non-positive. Hence,

$$\mathbb{E}[B_{x,y}B_{x,z} - D_{x,y}D_{x,z}] = 0, \qquad \mathbb{E}[B_{x,y}B_{x',y'} - D_{x,y}D_{x',y'}] = 0. \quad (7)$$

Reconstruction of Gordon's Theorem. We can recover the Gordon's conditions we assume during the last lecture, rewritten in the language of $B_{x,y}$ and $D_{x,y}$:

$$\begin{cases} \mathbb{E}[B_{x,y}^2] = (|x|^2\,|y|^2 + |x|^2\,|y|^2) = \mathbb{E}[D_{x,y}^2], \\ \mathbb{E}[B_{x,y}B_{x,z} - D_{x,y}D_{x,z}] = 0, \\ \mathbb{E}[B_{x,y}B_{x',y'} - D_{x,y}D_{x',y'}] \geq 0. \end{cases} \quad (8)$$

Accordingly, we can also rewrite the statement of Gordon's Theorem as

$$\Pr\left(\min_{x\in X}\max_{y\in Y}\left\{B_{x,y} + \psi(x,y)\right\} \geq c\right) = \Pr\left(\forall x, \exists y \ \text{ s.t. } D_{x,y} \geq c - \psi(x,y)\right), \quad (9)$$

where we denote $\lambda_{xy} = c - \psi(x,y)$.

2 Application - Linear Regression

One consequence of Theorem 1 is the following (see [Zhou-Koehler-Sutherland-Srebro '24] for the proof):

Theorem 2. *Let the data be generated as*

$$Y = Xw^\star + \xi, \tag{10}$$

where each row of $X \in \mathbb{R}^{n\times d}$ is drawn independently from $\mathcal{N}(0, I_d)$, and $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$ represents Gaussian noise.

Suppose there exists a function $F(w)$ such that for any w, with probability at least $1 - o(1)$,

$$\langle w - w^\star, x\rangle \le F(w). \tag{11}$$

Then, with high probability,

$$\sigma^2 + |w - w^\star|^2 \ \le\ \big(1 + o(1)\big)\left(\frac{1}{\sqrt{n}}\,|Y - Xw| + \frac{F(w)}{\sqrt{n}}\right)^2. \tag{12}$$

Remark 2. The inequality (12) can be interpreted as relating three types of errors:

- $\sigma^2 + \|w - w^\star\|^2$ — the *prediction error*;
- $|Y - Xw|/\sqrt{n}$ — the *training error*;
- $F(w)/\sqrt{n}$ — the *generalization error.*

2.1 Ordinary Least Squares (OLS)

Recall that the OLS optimum is defined as

$$\widehat{w}_{\text{OLS}} = \arg\min_{w\in\mathbb{R}^d} \|y - Xw\|_2. \tag{13}$$

From previous lectures, we have the following facts:

- $\frac{1}{n}\,|Y - X\widehat{w}_{\text{OLS}}|^2 \asymp \sigma^2(1 - \frac{p}{n})$.
- $\langle w - w^\star, X\rangle \le |w - w^\star|\cdot|X| \asymp |w - w^\star|\,\sqrt{d}$.

Here $|\cdot|$ denote the 2-norm, and we set $\gamma := \frac{p}{n}$.

For the OLS estimation, the inequality (12) becomes

$$\sigma^2 + |\widehat{w} - w^\star|^2 \ \le\ \left(\sigma\sqrt{1-\gamma} + |\widehat{w} - w^\star|\,\sqrt{\gamma}\right)^2, \quad \widehat{w} = \widehat{w}_{\text{OLS}}. \tag{14}$$

We now regard inequality (14) as a quadrature w.r.t. $r := |\widehat{w} - w^\star|$. Expanding the RHS and simplifying yields

$$r^2 - 2\sigma\sqrt{\tfrac{\gamma}{1-\gamma}}\,r + \sigma^2\tfrac{\gamma}{1-\gamma} \le 0. \tag{15}$$

Optimal Distance. Given the model $Y = Xw^\star + \xi$, we may rewrite the residual for any candidate parameter w as

$$Y - Xw = X(w^\star - w) + \xi, \tag{16}$$

where w is our approximation and ξ is the Gaussian noise. This decomposition provides the following interpretation for a near-optimal choice of w:

- The first term $X(w^\star - w)$ ensures that w does not stray too far from the true parameter $w^\star$.
- The Gaussian noise term ξ prevents w from being too close to $w^\star$. Indeed, in the special case $w = w^\star$, the residual consists solely of noise, with $\frac{1}{n}\,|Y - Xw|^2 \asymp \sigma^2(1-\gamma)$, which is undesirable as well.

3 CGMT

Theorem 3 (Von Neumann's Min - Max (informal))**.** *Let X, Y be convex sets, and let $f(x,y)$ be convex in x and concave in y. Then (under some additional assumptions)*

$$\min_{x\in X}\max_{y\in Y} f(x,y) = \max_{y\in Y}\min_{x\in X} f(x,y). \tag{17}$$

Remark 3. This equality characterizes the equilibrium of a *zero-sum game.*

Example 1. For instance, if $f(x,y) = \langle x, My\rangle$, then the equality follows directly from linearity, which is the saddle point of f.

Corollary 1 (Convex Gaussian Min–Max (Informal))**.** *If X and Y are convex sets, and the function $\psi(x,y)$ is convex in x and concave in y, then*

$$\min_{x\in X}\max_{y\in Y} PO(x,y) \approx \min_{x\in X}\max_{y\in Y} AO(x,y), \tag{18}$$

where PO and AO denote the Primary and Auxiliary Optimization problems, respectively.

The formal statement is in terms of tail probabilities as in the previous statement of GMT.

Remark 4. The direction

$$\min_x\max_y PO(x,y) \geq \min_x\max_y AO(x,y) \tag{19}$$

always holds, even for nonconvex settings.

A switching technique via convex - concave symmetry. For any function $f(x,y)$ that is concave in x and convex in y, we have

$$(-1)\cdot\min_x\max_y f(x,y) \;=\; \max_y\min_x\big(-f(x,y)\big), \tag{20}$$

with $-f$ being convex in x and concave in y. Moreover, by Von Neumann's min-max theorem,

$$\min_x\max_y\big(-f(x,y)\big) \;=\; \max_y\min_x\big(-f(x,y)\big), \tag{21}$$

which implies

$$(-1)\cdot\min_x\max_y f(x,y) \;=\; \min_y\max_x\big(-f(x,y)\big), \tag{22}$$

i.e.

$$\min_x\max_y f(x,y) \;=\; \max_y\min_x f(x,y). \tag{23}$$

Thus the convex-concave structure guarantees the interchangeability of the min and max operators, a key ingredient in proving the reverse direction in Corollary 1 (see Thrampoulidis-Oymak-Hassibi '15).