Chapter 9: Decomposability and Restricted Convexity:

9.1 Equation 9.10: $\Phi_{over}(\Theta) := \inf_{\substack{\Theta = \sum_j W_j \\ g \in G \\ W_g, g \in G}} \left\{ \sum_{g \in G} \|W_g\| \right\}$

"Norm" requirements:

1) Positive Homogeneity: $F(sX_1, \ldots, sX_n) = s^k F(X_1, \ldots, X_n)$

Proof: $\phi(s\theta) = s^k \phi(\theta)$

2) Positive Definite: $F \in \mathbb{R}^n$; $F(0) = 0$ and $F(X) > 0$ For $X \neq 0$

Proof: $\phi(0) = \inf(0) = 0$; $\phi(\theta) = \inf\{\theta\}$ for $\theta \neq 0$

3) Optimal Decomposition: $f(a+b) \leq f(a) + f(b)$

Proof: $\phi(\theta) = \sum_j \|W_g\|$

$\phi(\theta') = \sum_j \|W_g'\|$

$\phi(\theta + \theta') = \sum_j \|W_g + W_g'\|$

$\leq \sum_j \|W_g\| + \|W_g'\|$

$\leq \phi(\theta) + \phi(\theta')$

9.2. Equation 9.20: $\Theta_S := \underset{\tilde{\theta} \in S}{\arg\min} \|\tilde{\theta} - \theta\|^2$

a) $S \subseteq \{1, \ldots, d\}$

$M(S) := \{\theta \in \mathbb{R}^d \mid \theta_j = 0 \text{ for all } j \notin S\}$

$\theta := \underset{\theta \notin S}{\arg\min} \|\tilde{\theta} - \theta_j\|^2$

$= \sqrt{\sum_{j=1}^{d} (\tilde{\theta} - \theta_j)^2}$

$= \sqrt{\sum_{j=1}^{d} \tilde{\theta}^2}$

b. $M(U,V) := \{\theta \in \mathbb{R}^{d\times d} \mid \text{rowspan}(\theta) \subseteq U, \text{colspan}(\theta) \subseteq V\}$

$$\Theta := \underset{\text{rowspan}(\theta)\subseteq U}{\text{argmin}} \|\tilde{\theta} - \theta_j\|^2$$

$$= \underset{\text{colspan}(\theta)\subseteq V}{\text{argmin}} \|\tilde{\theta} - \theta_j\|^2$$

$$= \sqrt{\sum_{j=1}^{d} (\tilde{\theta} - \text{colspan}(\theta_j))^2} = \sqrt{\left(\begin{bmatrix}\tilde{\theta}\end{bmatrix} - \begin{bmatrix}\theta_1\end{bmatrix}\right)^2 + \left(\begin{bmatrix}\tilde{\theta}\end{bmatrix} - \begin{bmatrix}\theta\end{bmatrix}\right)^2 + \cdots + \left(\begin{bmatrix}\tilde{\theta}\end{bmatrix} - \begin{bmatrix}\theta_d\end{bmatrix}\right)^2}$$

$$= \sqrt{\sum_{j=1}^{d} (\tilde{\theta} - \text{rowspan}(\theta_j))^2}$$

$$= \sqrt{([\ \tilde{\theta}\ ] - [\ \theta_1\ ])^2 + ([\ \tilde{\theta}\ ] - [\ \theta_2\ ])^2 + \cdots + ([\ \tilde{\theta}\ ] - [\ \theta_d\ ])^2}$$

9.3. Equation 9.5:

$$\mathbb{P}_\theta(y|x) = h_\sigma(y) \cdot \exp\left\{\frac{y(\langle x, \theta^*\rangle) - \psi(\langle x, \theta^*\rangle)}{c(\sigma)}\right\}$$

a) $y = \langle x, \theta\rangle + w$ where $w \sim N(0, \sigma^2)$

$c(\sigma) = \sigma^2$; $\psi(\langle x, \theta^*\rangle) = \psi(t) = t^2/2$

$$P(y \mid t, \sigma^2) = N(y \mid t, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-t)^2}{2\sigma^2}}$$

$$= h_\sigma(y) \cdot \exp\left\{\frac{-\langle x, \theta\rangle + 2\langle x, \theta\rangle t - t^2}{2c(\sigma)}\right\}$$

$$= h_\sigma(y) \cdot \exp\left\{\frac{y(\langle x, \theta\rangle) - \psi(\langle x, \theta^*\rangle)}{c(\sigma)}\right\}$$

Where $h_\sigma(y) = 1/\sqrt{2\pi\sigma^2}$

b. Poisson Distribution:

$$P(\lambda, x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(\lambda, y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$$= \frac{e^{\langle x, \theta \rangle} \cdot e^{-t}}{y!} \quad \text{when } \lambda = e^{\langle x, \theta \rangle} = t$$

$$= \log\left[h_\theta(y) \cdot \exp\left\{\frac{\langle x, \theta \rangle - t}{c(\sigma)}\right\}\right]$$

$$P(\lambda, y) = \log P(y \mid x)$$

9.4

a) Dual Norm: l_1-norm is l_∞-norm:

$$l_1\text{-norm}: \|x\|_1 = \sum_{i=1}^{n} x_i = \max_{\|z\| \le 1} z^T x$$

$$l_\infty\text{-norm}: \|x\| = \sqrt[\infty]{\sum_{i=1}^{n} x_i^\infty} = \max_{\|z\| \le 1} z^T x$$

b) Dual Norm: l_p-norm and l_q-norm:

$$l_p\text{-norm}: \|u\|_p = \sqrt[p]{\sum_{i=1}^{n} (u_i)^{1/p}}$$

$$l_q\text{-norm}: \|v\|_q = \sqrt[q]{\sum_{i=1}^{n} (v_i)^{1/q}}$$

These relate through an inequality: $\phi(uv) \le \phi(u)\,\phi(v)$.

$$\phi(uv) = \sum_{i=1}^{n} uv$$

$$\le \left(\sum_{i=1}^{n} (u_i)^p\right)^{1/p} \cdot \left(\sum_{i=1}^{n} (v_i)^q\right)^{1/q}$$

$$\le \phi(u) \cdot \phi(v)$$

where $1 = \frac{1}{p} + \frac{1}{q}$

c) Dual Norm: Nuclear Norm is the l_2-operator norm:

$$\text{Nuclear Norm: } \phi^*(N) = \sup_{\|Z\|_2=1} \|NZ\|_2 \le \sqrt{\left(\sum_{i=1}^{N} N_i^2\right)\left(\sum_{i=1}^{N} z^2\right)}$$

$$\le \sqrt{\sum_{j=1}^{N} N_i^2}$$

$$l_2\text{-norm: } \|N\|_2 = \sqrt{\sum_{i=1}^{N} N_i^2}$$

9.5 (Overlapping Group Norm)

$$\phi_{over}(\theta) = \inf_{\substack{\theta=\sum_{g\in G} w_g \\ w_g, g\in G}} \left\{ \sum_{g\in G} \|w_g\| \right\} = \|N\|_2$$

a) $$\phi(v,u) = \sum_{i=1}^{N} v_i \circ u_i$$

$$= \max_{v\in\mathbb{R}^d} \langle v, u\rangle \quad \text{such that } \|v\|_2 \le 1 \text{ for all } g\in G$$

b) By "Hölder's Inequality",

$$\phi(v,u) = \sum_{i=1}^{N} v_i u_i$$

$$= (v_1u_1 + v_2u_2 + \circ\circ\circ + v_N u_N)$$

$$\le \sqrt[q]{(v_1 + v_2 + \circ\circ\circ + v_N)^q} \circ \sqrt[p]{(u_1 + u_2 + \circ\circ\circ + u_N)^p}$$

$$\le \phi(v) \circ \phi(u)$$

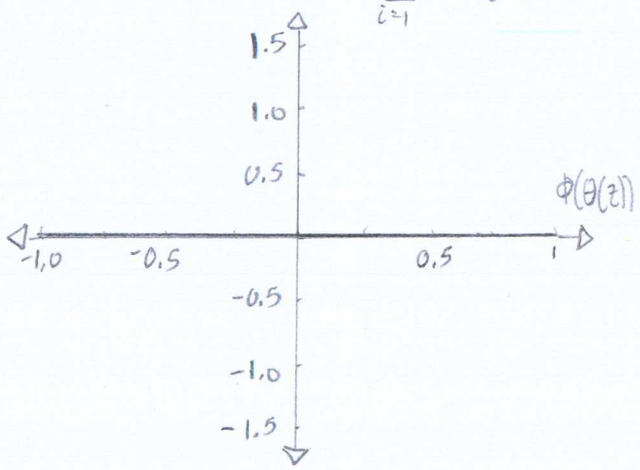
$$\le \max_{g\in G} \|v_g\|_p \circ \max_{g\in G} \|u_g\|_q$$

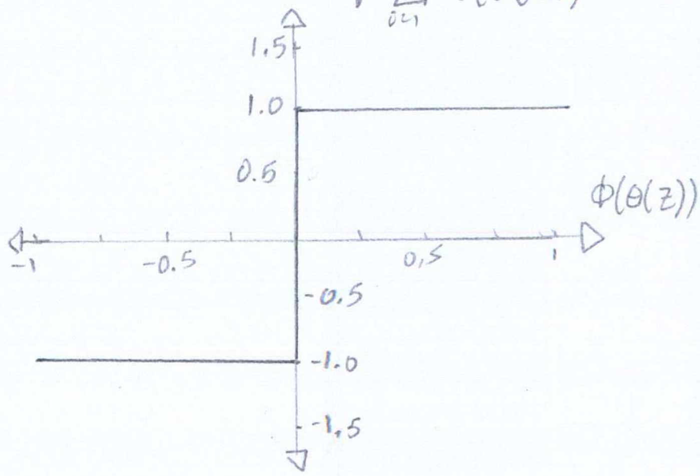
9.6 $\phi: \mathbb{R}^d \to \mathbb{R}$ be a norm

$\theta \in \mathbb{R}^d$ be arbitrary

$$\frac{d}{dz}\phi(\theta) = \frac{d}{dz}\sqrt{\sum \phi(\theta(z))^2} = \frac{\phi(\theta(z))}{\sqrt{\sum \phi(\theta(z))^2}};$$

$$\inf \frac{d\Phi(\theta(z))}{dz} = \frac{\Phi(\theta(z))}{\sqrt{\sum_{i=1}^{\infty} \Phi(\theta(z))^2}} \qquad \sup \frac{d\Phi(\theta(z))}{dz} = \frac{\Phi(\theta(z))}{\sqrt{\sum_{i=1}^{n} \Phi(\theta(z))}}$$





9.7.

a) $$|\langle u, v\rangle| = \sum_{i=1}^{N} |u_i \circ v_i|$$

$$= |(u_1 v_1 + u_2 v_2 + \circ\circ\circ + u_N v_N)|$$

$$\leq |(u_1 + u_2 + \circ\circ\circ + u_N)||(v_1 + v_2 + \circ\circ\circ + v_N)|$$

$$\leq \sum_{i=1}^{N} |u_i| \cdot \sum_{i=1}^{N} |v_i|$$

$$\leq \Phi(u) \circ \Phi(v)$$

b) $$|\langle u, v\rangle| = \sum_{i=1}^{N} |u_i v_i|$$

$$= |(u_1 v_1 + u_2 v_2 + \circ\circ\circ + u_N v_N)|.$$

$$\leq |(u_1 + u_2 + \circ\circ\circ + u_N)||(v_1 + v_2 + \circ\circ\circ + v_N)|$$

$$\leq \sqrt[p]{|(u_1 + u_2 + \circ\circ\circ + u_n)^p|} \circ \sqrt[q]{|(v_1 + v_2 + \circ\circ\circ + v_n)^q}$$

$$\leq \|u\|_p \|v\|_q$$

c) $$|\langle u, v\rangle| \leq \|u\|_p \|v\|_q$$

$$\leq Q \|u\|_p \|v\|_q$$

$$\leq Q^{\frac{1}{p}} \cdot Q^{\frac{1}{q}} \|u\|_p \|v\|_q$$

$$\leq \|uQ\|_p \|vQ\|_q$$

9.8. Equation (9.4.1) $\mu_n(\phi^*) = \mathbb{E}_{X,\varepsilon}\left[\phi^*\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i\right)\right]$

a) $\phi(\theta) = \sum \|\theta_g\|$

$\varepsilon_i \geq \frac{\|X\Delta_2^2\|}{2n} = \sigma^2$ for a sub-gaussian

$$\mu(\phi^*) = \mathbb{E}\left[\phi^*\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i\right)\right]$$

$$= \mathbb{E}\left[\phi^*\left(\frac{1}{n}\sum_{i=1}^{n}\sigma^2 X_i\right)\right]$$

$$= \mathbb{E}\left[\sigma\,\phi^*\left(\frac{1}{n}\sum\|\theta_g\|\right)\right]$$

$$= \mathbb{E}\left[\sigma\,\phi^*\left(\frac{1}{n}\sum\|\theta_m\| + \frac{1}{n}\sum\|\theta_G\|\right)\right]$$

$$= \sigma\sqrt{\frac{m}{n}} + \sigma\sqrt{\frac{\log G}{n}}$$

b) $\mu(\phi^*) = \mathbb{E}\left[\phi^*\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i\right)\right]$

$$= \mathbb{E}\left[\phi^*\left(\frac{1}{n}\sum\sigma^2 X_i\right)\right]$$

$$= \mathbb{E}\left[\sigma\,\phi^*\left(\frac{1}{n}\sum\|\theta_g\|\right)\right]$$

$$= \mathbb{E}\left[\sigma\,\phi^*\left(\frac{1}{n}\sum\|\theta_{d_1}\| + \frac{1}{n}\sum\|\theta_{d_2}\|\right)\right]$$

$$= \sigma\sqrt{\frac{d_1}{n}} + \sigma\sqrt{\frac{d_2}{n}}$$

9.9. $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{k}{2}\|y-x\|_2^2$ for all $x, y \in \mathbb{R}^d$

$$\langle \nabla f(y) - \nabla f(x), y-x \rangle = \|(\nabla f(y) - \nabla f(x))\|\,\|y-x\|$$

$$= \|\nabla f(y)\|\,\|y-x\| - \|\nabla f(x)\|\,\|y-x\|$$

$$= f(y) - f(x) - \|\nabla f(x)\|\,\|y-x\|$$

$$\geq \frac{k}{2}\|y-x\|_2^2$$

$$\boxed{\begin{array}{l} \nabla f(y) = \dfrac{f(y)-f(x)}{y-x} \\ \nabla f(y)(y-x) = f(y) - f(x) \end{array}}$$

Convexity: 1) No assumptions (Zeroth-order)

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

2) Differentiable (First-order)

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

3) Twice Differentiable (Second-order)

$$\nabla^2 f(x) \geq 0$$

General rules of thumb about convex functions.

9.10. $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{k}{2} \|y - x\|_2^2$

$\langle \nabla f(y) - \nabla f(x), y - x \rangle = \|\nabla f(y) - \nabla f(x)\| \, \|y - x\|$

Slope:

$\nabla f(y) = \dfrac{f(y) - f(x)}{y - x}$

$\nabla f(y)(y-x) = f(y) - f(x)$

$= \|\nabla f(y)\| \|y - x\| - \|\nabla f(y)\| \|y - x\|$

$= \nabla f(y) - \nabla f(x) - \|\nabla f(y)\| \|y - x\|$

$\geq \frac{k}{2} \|y - x\|_2^2$

(Equation 9.56) Dual-Norm: ℓ_∞-norm Restricted Curvature

$$\|\hat{\Sigma} \Delta\|_\infty \geq k \|\Delta\|_\infty - \tau \|\Delta\|_1 \quad \text{for all } \Delta \in \mathbb{R}^d$$

(Equation 9.57) ℓ_∞-norm Restricted Eigenvalues

$$\|\hat{\Sigma} \Delta\|_\infty \geq k \|\Delta\|_\infty \qquad \text{for all } \Delta \in C(S; \alpha)$$

"These equations set a lower bound on variance in an infinite data set, per se on infinite cardinality in the set."

9.11. $\tau_n = c_1\sqrt{\frac{\log d}{n}}$ $\qquad n > c_2|S|^2 \log d$

$k' = k/2$ $\qquad c_2 = \frac{4c_1^2(1+\alpha)^4}{k^2}$

"Many of these derive from chapter 7's sparse data analysis with relations to data set 'size'."

From equation 9.56, $\|\hat{\Sigma}\Delta\|_\infty \geq k\|\Delta\|_\infty - \tau\|\Delta\|_1$

$$\geq k\|\Delta\|_\infty - c_1\sqrt{\frac{\log d}{n}}\|\Delta\|_1$$

$$\geq k\|\Delta\|_\infty - \sqrt{\frac{k^2 c_2 \log d}{4c_1^2(1+\alpha)^4 c_2 |S|^2 \log d}}\|\Delta\|_1$$

$$\geq k'\left(\|\Delta\|_\infty - \frac{\|\Delta\|_1}{c_1(1+\alpha)^2|S|^2}\right)$$

When $|S|$, data size is large

$$\geq k'\|\Delta\|_\infty$$

9.12 $y = X\theta^* + w$

a) Equation 7.6b $T_\lambda(y) = \begin{cases} \text{sign}(y_i)(|y_i| - \lambda) & \text{if } |y_i| > \lambda \\ 0 & \text{otherwise} \end{cases}$

Derived in Problem 7.2

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \left\{\frac{1}{2}\|\theta\|_2^2 - \langle\theta, \tfrac{1}{n}X^T y\rangle + \lambda_n\|\theta\|_1\right\}$$

$$= \theta - \frac{1}{n}X^T y + \lambda_n$$

$$\lambda^* = \frac{1}{n}X^T y - \theta$$

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \left\{\frac{1}{2}\|\theta\|_2^2 - \langle\theta, \tfrac{1}{n}X^T y\rangle + \left(\tfrac{1}{n}X^T y - \theta\right)\|\theta\|_1\right\}$$

$= (\theta+1)\frac{1}{n}X^T y$ @ $\theta = 1$; $\hat{\theta} = T_\lambda(y)\frac{1}{n}X^T y$

b) $\lambda_n \geq 2\{\|(\frac{X^TX}{n} - I_d)\theta^*\|_\infty + \|\frac{X^Tw}{n}\|_\infty\}$

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \{\tfrac{1}{n}\|\theta\|_2^2 - \langle \theta, \tfrac{1}{n}X^Ty\rangle + \lambda_n\|\theta\|_1\}$$

$$= \theta - \tfrac{1}{n}X^Ty + \lambda$$

$$= \theta - \tfrac{1}{n}X^T(X^T\theta^* + w) + 2\{\|(\tfrac{X^TX}{n} - I_d)\theta^*\|_\infty + \|\tfrac{X^Tw}{n}\|_\infty\}$$

$$= \underbrace{\theta - \tfrac{1}{n}X^TX\theta^*}_{\text{"one count"}} + \tfrac{1}{n}X^Tw + \underbrace{2\{\|(\tfrac{X^TX}{n} - I_d)\theta^*\|_\infty + \|\tfrac{X^Tw}{n}\|_\infty\}}_{\text{"two counts"}}$$

"one count" $+$ "two counts" $=$ "three counts"

$$= 3\sqrt{s}\,\|\tfrac{X^TX}{n} - 1\|\theta$$

$$\geq \tfrac{3}{2}\sqrt{s}\,\lambda_n$$

c) $\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{2}\sqrt{s}\,\lambda_n$

$$\leq 3\sqrt{s}\,\underbrace{(v(v\|\theta^*\|_2 + b)\{\sqrt{\tfrac{\log d}{n}} + \delta\})}_{\frac{\lambda_n}{2}}$$

$$\lambda = 2v(v\|\theta\|_2 + b)\{\sqrt{\tfrac{\log d}{n}} + \delta\}$$

(Corallary 9.27) Generalized Linear Model Lasso

$$\|\hat{\theta} - \theta^*\|_\infty \leq 3\frac{\lambda_n}{k} \qquad \text{when} \qquad \lambda = 2Bc\left(\sqrt{\tfrac{\log d}{n}} + \delta\right)$$

"Maximum variance in an infinite dimension data set when the number of points in a single dimension satisfies $n > c_0^2 s \log d$."

9.13

$$\|\hat{\theta}-\theta^*\|_1 \le \frac{3\lambda_n}{k} \qquad \|\hat{\theta}-\theta^*\|_2 \le \frac{3\cdot\lambda_n}{k}$$

$$\lambda = 2BC\sqrt{\frac{\log d}{n}} \qquad \lambda = 2BC\sqrt{\frac{\log d}{n}}$$

$$B = 4s \qquad B = 2\sqrt{s}$$

$$C = \sigma \qquad C = \sigma$$

9.14. $Z = (X, y) \in \mathcal{X} \times \mathcal{Y}$

$$\left|\frac{\partial L(\theta; z)}{\theta_j} - \frac{\partial L(\tilde{\theta}; z)}{\theta_j}\right| \le L\,|X_{ij}\langle x_i, \theta-\tilde{\theta}\rangle|$$

a) $\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{\psi(\langle x_i, \theta\rangle) - y_i\langle X_i, \theta\rangle\}$

$$\nabla\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{\psi'(\langle x_i, \theta^*\rangle) - y_i\}x_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} V_i$$

$$\log\mathbb{E}[e^{-tV_{ij}}] = \log\mathbb{E}[e^{ty_i x_i}] - tx_{ij}\psi'(\langle x_i, \theta^*\rangle)$$

$$= \psi(tx_{ij} + \langle x_i, \theta^*\rangle) - \psi(\langle x_i, \theta^*\rangle) - tx_{ij}\psi'(\langle x_i, \theta^*\rangle)$$

$$= \frac{1}{2}t^2 x_{ij}^2\psi''(\tilde{t}x_{ij} + \langle x_i, \theta^*\rangle)$$

$$\le \frac{B^2}{2}t^2 x_{ij}^2$$

$$\frac{1}{n}\log\mathbb{E}[e^{-t\sum V_{ij}}] \le \frac{B^2}{2}\left[\frac{t^2}{n}\sum_{i=1}^{n} x_{ij}^2\right]$$

$$\le L\,|X_{ij}\langle x_i, \theta-\tilde{\theta}\rangle| \quad \text{where } L = \frac{B^2}{2}$$

b) $\mathbb{T}(R_{j\rho}) = \{\Delta \in \mathbb{R}^d \mid \frac{\|\Delta\|_1}{\|\Delta\|_\infty} \le \rho, \text{ and } \|\Delta\|_\infty \le r\}$

$$V_j = \frac{1}{4Lr\rho}\sup_{\Delta\in\mathbb{T}}\left|\frac{1}{n}\sum_{i=1}^{n} f_j(\Delta, z)\right| \quad \text{for } j = 1, \ldots, d$$

$$f(\Delta; z_j) = \left\{\frac{\partial\mathcal{L}(\theta^*+\Delta; z_j)}{\partial\theta_j} - \frac{\partial\mathcal{L}(\theta^*; z)}{\partial\theta_j}\right\} - \left\{\frac{\partial\bar{\mathcal{L}}(\theta^*+\Delta)}{\partial\theta_j} - \frac{\partial\bar{\mathcal{L}}(\theta^*)}{\partial\theta_j}\right\}$$

$$\mathbb{E}_x\left[e^{\lambda \|V\|_\infty}\right] \stackrel{\circ}{=} \mathbb{P}\left[|\lambda \|V\|_\infty| > \lambda x\right]$$

$$\stackrel{\circ}{=} \mathbb{P}\left[e^{\lambda \||V\||_\infty} > e^{\lambda x}\right]$$

$$\leq \exp\left\{\frac{-nt^2}{2B^2C^2} + \log d\right\} \quad \text{from Equation 9.63A}$$

$$\leq d \circ \exp\left(\lambda \left\|\frac{1}{n}\sum_i \varepsilon_i x_{ij}^2\right\|\right) \text{ where } t = \frac{x_{ij}}{n}$$

$$\varepsilon = (B^2C^2)^2$$

$$\lambda = \frac{1}{2}$$

c) $$\mathbb{P}\left[\|V\|_\infty \geq t\right] = \mathbb{P}\left[\|\nabla \mathcal{L}_n(\theta^*)\|_\infty \geq t\right]$$

$$\stackrel{\circ}{=} \mathbb{P}\left[\lambda \|\nabla \mathcal{L}(\theta^*)\|_\infty \geq \lambda t\right]$$

$$\stackrel{\circ}{=} \mathbb{P}\left[e^{\lambda \|\nabla \mathcal{L}(\theta^*)\|_\infty} \geq e^{\lambda t}\right]$$

$$= \mathbb{P}\left[e^{1 + \lambda \mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|] + \frac{\lambda^2 \mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|^2]^2}{2} + \cdots} \geq e^{\lambda t}\right]$$

$$\leq \mathbb{P}\left[e^{\frac{\lambda^2 \mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|^2]^2}{2}} \geq e^{\lambda t}\right]$$

$$\underset{\lambda}{\operatorname{argmin}}\left\{\frac{\lambda^2 \mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|]^2}{2} - \lambda t\right\} = 0$$

$$\lambda^* = \frac{t}{\mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|]^2}$$

$$\mathbb{P}\left[e^{\frac{\lambda^2 \mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|]^2}{2}} \geq e^{\lambda t}\right] = e^{\frac{t^2 \mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|]^2}{2\mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|^2]^4} - \frac{t^2}{\mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|^2]^2}}$$

$$= e^{\frac{-t^2}{2\mathbb{E}[\|\nabla \mathcal{L}(\theta^*)\|^2]^2}}$$

$$= e^{\frac{-t^2}{2\sigma^2}}$$

<u>One-sided:</u> $e^{-\frac{t^2}{2\sigma^2}}$ <u>Two-sided:</u> $2 \cdot e^{-\frac{t^2}{2\sigma^2} + \log d}$

Note: No absolute signs around the norm, $\|V\|_\infty$, suggests an abstraction about one-, or two-sided.

$$\mathbb{P}\left[\|V\|_\infty \geq t\right] \stackrel{\circ}{=} e^{\frac{-t^2}{2\sigma^2}} \longrightarrow 2 \cdot e^{-\frac{t^2}{2\sigma^2} + \log d} = 2de^{-\frac{t^2}{2\sigma^2}}$$

d) $\|\nabla \tilde{\mathcal{L}}(\theta^*+\Delta) - \nabla \mathcal{L}(\theta^*)\|_\infty \geq k\|\Delta\|_\infty$ for all $\Delta \in \Pi(r;\rho)$

$$\|\nabla \mathcal{L}_n(\theta^*+\Delta) - \nabla \mathcal{L}_n(\theta^*)\|_\infty \geq \underbrace{\|\nabla \tilde{\mathcal{L}}(\theta^*+\Delta) + \nabla \mathcal{L}(\theta^*)\|_\infty}_{\text{"Estimation Error"}} - \underbrace{\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle}_{\text{"Approximate Error"}}$$

$$\geq k\|\Delta\|_\infty - \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle$$

$$\geq k\|\Delta\|_\infty - |\mathcal{E}(\Delta) - \mathcal{E}(\hat{\Delta})|$$

$$\geq k\|\Delta\|_\infty - 16 L \Phi(\Delta)\delta \text{ (Theorem 9.34)}$$

A set $(\Delta \in \Pi(r,\rho))$ bounds approximate error and the second term within the inequality, $\Phi(\Delta)$.

The guess is $\Delta \in \Pi(r,\rho) \in B_2(r) \cap \{\Phi(\Delta) \leq \rho\|\Delta\|_2\}$

$$\in \{\Delta \in \mathbb{R}^d, \sigma^2\sqrt{\tfrac{\log d}{n}}\tfrac{r}{\delta} \leq \|\Delta\|_2\} \cap \{\Phi(\Delta) \leq \rho\|\Delta\|_2\}$$

$$\|\nabla \mathcal{L}_n(\theta^*+\Delta) - \nabla \mathcal{L}_n(\theta^*)\|_\infty \geq k\|\Delta\|_\infty - 16 L \Phi(\Delta)\delta$$

$$\geq k\|\Delta\|_\infty - 16 L \frac{\Phi(\Delta)}{\|\Delta_2\|} \|\Delta\|_2 \circ \delta$$

$$\geq k\|\Delta\|_\infty - 16 L \circ \rho \circ \sigma^2\sqrt{\frac{\log d}{n}}\frac{r}{\delta} \circ \delta$$

$$\geq k\|\Delta\|_\infty - 16 L \sigma^2\sqrt{\frac{\log d}{n}}\rho r$$