Old tensor mle stuff

Cole Franks, Rafael Oliveira, Akshay Ramachandran, Michael Walter

February 2020

Contents

1 Old stuff

1.1 Old lower bound lemma

In the conclusion of the lemma we needed to use that convergence in total variation of some estimator $\widehat{\Theta}_n$ to another, $\widehat{\Theta}$, implied that the former has minimax error at least that of the latter in any dissimilarity measure. This holds by applying the next lemma to the random variables $X_n = d_*(\widehat{\Theta}_n, \Theta)$ and $Y = d_*(\widehat{\Theta}, \Theta)$ where d_* represents any nonnegative function. For example, we could take d_* to be the Frobenius, spectral, Fischer-Rao, Kullback-Leibler or Mahalanobis "distances". [CF: surely I can cite this next thing, I am just proving it for my own sanity]

Lemma 1. *Suppose X_n, Y are nonnegative random variables such that $X_n \to Y$ in d_{TV}. Then*

$$\limsup_{n\to\infty} \mathbb{E}X_n \geq \mathbb{E}Y.$$

Proof. If the mean of Y is bounded then we have Markov's inequality. Let $\varepsilon > 0$; by the Dominated Convergence Theorem there is α large enough that

$\mathbb{E}[Y 1_{Y\le\alpha}] \ge \mathbb{E}[Y] - \varepsilon$. Now we have

$$\mathbb{E}[X_n] \ge \mathbb{E}[X_n 1_{X_n\le\alpha}] \to \mathbb{E}[Y 1_{Y\le\alpha}] \ge \mathbb{E}[Y] - \varepsilon.$$

as $n \to \infty$, where the limit is deduced by Hölder's inequality. Letting $\varepsilon \to 0$ completes the proof. □

1.2 Different Inner Product

Definition 1. *For desired marginals $\{R_a^2\}_{a\in[k]}$ (assume for now R are Hermitian though we can pick different square roots if required), define inner product*

$$\langle Z, Y\rangle_R := \sum_a \langle R_a Z R_a^*, Y\rangle$$

$$\|Z\|_R^2 := \langle Z, Z\rangle_R = \sum_a \|R_a Z_a\|_F^2$$

We restate the projective likelihood function and define gradient and Hessian in this metric:

Definition 2.

$$f_{\boldsymbol{X}}(\Theta_1, \ldots, \Theta_n) = \log \left\langle \sum_{i\in[n]} X_i X_i^*, \bigotimes_{a\in[k]} \Theta_a \right\rangle - \sum_{a\in[k]} \frac{1}{d_a} \log\det \Theta_a$$

Also $\rho := \sum_i X_i X_i^*$ and $\{\rho^S\}_{S\subseteq[k]}$ are marginals.

Fact 2.

$$(\nabla f(I))_a = d_a \rho^{\{a\}} - I_a$$

Proof. We can define ∇f dually as $\forall Z : \langle \nabla f(I), Z\rangle_R := \partial_{t=0} f(e^{tZ})$

$$\partial_{t=0} f(e^{tZ_a}) = \partial_{t=0} \langle \rho, I_{\overline{a}} \otimes e^{tZ_a}\rangle - \partial_{t=0} \frac{1}{d_a} \log\det e^{tZ_a}$$

$$= \left\langle \rho^{\{a\}} - \frac{1}{d_a} I_a, Z_a e^{tZ_a} \right\rangle |_{t=0} = \left\langle R_a^{-1} \left(\rho^{\{a\}} - \frac{1}{d_a} I_a \right) R_a^{-1}, Z_a \right\rangle_R$$

Similarly we define the Hessian as

$$\partial_{s=t=0} f(e^{tZ_a + sY_b}) = \langle \rho, \{I_{\overline{a}} \otimes Z_a, I_{\overline{b}} \otimes Y_b\}\rangle$$

$$\implies (\nabla^2 f(I))_{aa} = \langle R_a^{-1} \rho^{\{a\}} R_a^{-1}, \{Z, Y\}\rangle_R$$

$$\implies (\nabla^2 f(I))_{ab} = \langle \rho^{\{a,b\}}, Z \otimes Y\rangle$$

□

[AR: Not sure how to define Hessian. I think I'd like the Hessian to be

$$\sum_a E_{aa} \otimes (1 \pm \epsilon) I + \sum_{a \neq b} E_{ab} \otimes \pm\lambda$$

for some small ϵ, λ. Then the Hessian will be $1 - \epsilon - (k-1)\lambda$-strongly convex.]

Lemma 3 (Restatement of **??**)**.** *Let f be geodesically convex everywhere. All the below quantities are wrt metric $\langle \cdot, \cdot \rangle_R$. Assume f and λ-strongly geodesically convex ball of radius κ about I; further assume the geodesic gradient satisfies $\|\nabla f(I)\|_R = \varepsilon < \lambda\kappa$. Then there is an optimizer within an ε/λ-ball.*

Proof of **??**. The proof is exactly the same except the following:

$$g'(0) = \langle \nabla f(I), Z \rangle_R \geq -\|\nabla f(I)\|_R \|Z\|_R \geq -\varepsilon$$

□

Remark 1. *Note the perturbation lemma then gives the following strategy. By Cole's lemma, we have that $c\|\nabla f(I)\|_R \geq \|\nabla f(I)\|_{op}$. If we can say the same thing for the optimizer Z, then it is enough for $\lambda\kappa \geq \Omega(1/c) > \varepsilon$ and we can improve sample complexity to $nD > c \max_a d_a^2$.*

A similar thing is true if we can show the above inequality for the gradient flow for $\log \max_a d_a$ time.

Lemma 4. *λ-strong convexity is a sufficient condition for fast convergence of the gradient flow:*

$$-\partial_{t=0} \|\nabla f(e^{-t\nabla f(I)})\|_R^2 = -\partial_{t=0}^2 f(e^{-t\nabla f(I)}) = \langle \nabla^2 f, \nabla f \otimes \nabla f \rangle \geq \lambda \|\nabla f\|_R^2$$

[AR: Not sure how to write the third term above, the inner product with Hessian]

1.3 Old proof of ??

Proof of **??**. Take any quadratic form of the Hessian for $\{Z_a \perp I_a\}$:

$$\partial_{t=0}^2 f(e^{tZ}) = \sum_a \langle Q^a, Z_a^2 \rangle + \sum_{a \neq b} \langle Q^{ab}, Z_a \otimes Z_b \rangle$$

$$\geq \sum_a \lambda_{\min}(Q^a) \|Z_a\|_F^2 - \sum_{a \neq b} \|Q^{ab}(I - P_{ab})\|_{op} \|Z_a\|_F \|Z_b\|_F$$

Now we can use our high probability bounds derived above:

$$\forall a : Q^a \succeq \frac{1-\epsilon}{d_a} I_a; \qquad \forall a \neq b : \|Q^{ab}(I - P_{ab})\|_{op} < \frac{\lambda}{\sqrt{d_a d_b}}$$

$$\implies \partial_{t=0}^2 f(e^{tZ}) \geq \sum_a \frac{1-\varepsilon}{d_a} \|Z_a\|_F^2 - \sum_{a \neq b} \frac{\lambda}{\sqrt{d_a d_b}} \|Z_a\|_F \|Z_b\|_F$$

$$\geq \sum_a \frac{1-\varepsilon+\lambda}{d_a}\|Z_a\|_F^2 - \lambda\left(\sum_a \frac{1}{\sqrt{d_a}}\|Z_a\|_F\right)^2$$
$$\geq \sum_a \frac{1-\varepsilon+\lambda}{d_a}\|Z_a\|_F^2 - k\lambda\sum_a \frac{1}{d_a}\|Z_a\|_F^2$$
$$= (1-\varepsilon-(k-1)\lambda)\|Z\|^2$$

Choosing ε, λ small enough gives the theorem. □

Proof: *[CF: Akshay's conceptual proof of ??].* We can in fact show that ∇^2 is well-conditioned using the following:

$$-\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \preceq \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \preceq \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

[AR: Ya ∇_{ab} notation is fine, just needed something that was a matrix of the right dimensions, so shorthand M was to avoid weird things with ρ]
[AR: It's fine if they're of different sizes, we enumerate the basis of the whole space as $\cup_a e_a \otimes \{e_{i\in[d_a]}\}$]
[CF: $E_{aa} \otimes \nabla^2_{aa}$ is $kd_a^2 \times kd_a^2$ dimensional. So how does this make sense? Maybe needs to be updated along the lines of the next proof.]

$$\nabla^2 f = \sum_a E_{aa} \otimes \nabla^2_{aa} + \sum_{a\neq b} E_{ab} \otimes \nabla^2_{ab}$$

Now we can again use the high-probability bounds derived above: [TODO: actually cref them]

$$\nabla^2_{aa} \in \frac{1\pm\varepsilon}{d_a}; \qquad \forall a \neq b : \|\nabla^2_{ab}\|_{op} \leq \frac{\lambda}{\sqrt{d_a d_b}} \tag{1}$$

$$\nabla^2 \preceq \sum_a E_{aa} \otimes \left(\frac{1+\varepsilon}{d_a} I_a\right) + \sum_{a<b} E_{aa} \otimes \left(\frac{\lambda}{d_a} I_a\right) + E_{bb} \otimes \left(\frac{\lambda}{d_b} I_b\right)$$
$$\preceq \sum_a E_{aa} \otimes \frac{1+\varepsilon+(k-1)\lambda}{d_a} I_a$$

The same sequence of inequalities can be reversed to show a lower bound. So in fact we can show the above bounds on blocks shows $1+O(\varepsilon+k\lambda)$-condition number bound on the Hessian in norm $\|\cdot\|_d$. □

2 Old gradient bounds

[TODO: lower bound Hessian for operators and tensors for all different formats; we hope to get strong convexity with $\prod d_i/(d_1^2+\cdots+d_k^2)$ samples. I am concerned that a KLR19-style operator norm type theorem is needed to get $\tilde{O}$ of

this, but we will do what we can with the Frobenius bounds for now; I'd expect to need at least $\max_i \sqrt{d_i}$ too many samples.]
[TODO: It would also be nice to have that tight example for the log in KLR19...]

We recall the moment map and Hessian calculations

$$\partial_{t=0} f(e^{tX_a}) = \langle \nabla_a, X \rangle = \langle Q^a - sI_a, X \rangle$$

$$\partial^2_{t=0} f(e^{tX_a}) = \langle X, (\nabla^2)_{aa} X \rangle = \langle Q^a, X^2 \rangle$$

$$\partial_{s=0}\partial_{t=0} f(e^{tX_a} \otimes e^{sY_b}) = \langle Y, (\nabla^2)_{ab} X \rangle = \langle Q^{ab}, X \otimes Y \rangle$$

3 Operator Scaling

In this section we have n samples of $X \sim \mathcal{N}(0, \frac{1}{n}(\frac{1}{d_1}I_1) \otimes (\frac{1}{d_2}I_2))$. We will denote $D := d_1 d_2$. In order to use the KLR analysis, we will show that the one-body marginals have low error in $\|\cdot\|_{op}$ and the whole operator is a sufficient expander at the start.

3.1 Bernstein Proof of $\|\mu\|_{op}$

This is proven using matrix concentration

Theorem 5 (Bernstein)**.** *Consider independent $\{X_k\}$ such that $\mathbb{E}X_k = 0$ and $\lambda_{max}(X_k) \leq R$ almost surely. Further let the variance be $\sigma^2 := \|\sum_k \mathbb{E}X_k^2\|_{op}$.*

$$\begin{aligned} \mathbb{P}[\lambda_{max}\left(\sum_k X_k\right) \geq t] &\leq d \exp\left(-\frac{\Omega(t^2)}{\sigma^2 + tR}\right) \\ &\leq \begin{cases} d\exp(-\Omega(t^2/\sigma^2)) & \text{if } t \leq \sigma^2/R \\ d\exp(-\Omega(t/R)) & \text{if } t \geq \sigma^2/R \end{cases} \end{aligned}$$

In our setting, Q^a is comprised of $N := \frac{TD}{d_a}$ copies of a rank one gg^* where each Gaussian is $g \sim \mathcal{N}(0, N^{-1}\frac{1}{d_a}I_a) = \mathcal{N}(0, \frac{1}{TD}I_a)$. We will drop subscripts for d_a, I_a etc when they can be understood from context. Therefore we define $X := gg^* - \frac{1}{TD}I_a$ and note the following parameters:

$$\lambda_{max}(X) = \|g\|_2^2 - \frac{1}{TD} \qquad \lambda_{min}(X) = -\frac{1}{TD}$$

While $\|g\|_2$ is unbounded, we can threshold our distribution with a small loss in probability. Since we will be using χ^2 distributions much from now on, we will do a quick exercise to prove our threshold bounds:

Definition 3. *$\chi(\mu, d)$ denotes the χ^2 distribution with mean μ and d degrees of freedom. Explicitly $X \sim \chi(\mu, d) \implies X = \frac{\mu}{d}\sum_{i=1}^d g_i^2$ where $g \sim \mathcal{N}(0,1)$.*

Lemma 6. *For $X \sim \chi(\mu, d)$ we have the following (explicit and approximate) formula for the MGF, $\forall \theta < \left(O(\frac{\mu}{d})\right)^{-1}$:*

$$\begin{aligned}\log \mathbb{E} \exp(\theta X) &= -\frac{d}{2} \log\left(1 - 2\theta \frac{\mu}{d}\right) \\ &\leq \theta\mu + \theta^2 \frac{O(\mu^2)}{2d}\end{aligned}$$

Theorem 7 (Sub-exp variables)**.** *The above MGF bound gives tail decay:*

$$\forall \theta < b^{-1} : \log \mathbb{E} \exp(\theta(X - \mathbb{E}X)) \leq \theta^2 \frac{\sigma^2}{2}$$

$$\implies \mathbb{P}[X - \mu \geq t] \leq \begin{cases} \exp(-\Omega(t^2/\sigma^2)) & t \leq \sigma^2/b \\ \exp(-\Omega(t/b)) & t \geq \sigma^2/b \end{cases}$$

With these bounds in mind, note our variables $\|g\|_2^2 \sim \chi(\frac{d_a}{TD}, d_a)$ so we have $\sigma^2 = \frac{d}{(TD)^2}, b = \frac{1}{TD} \implies \sigma^2/b = \frac{d}{TD}$

$$\mathbb{P}[\exists k : \lambda_{max}(X_k) \geq M\sqrt{\log N} \frac{d}{TD}] \leq \exp(-\Omega(M^2))$$

If we're happy with $1/poly$ failure probability we will take $M^2 \sim \log D$, so in our matrix bound $R_{max} \leq \frac{d \log D}{TD}$

$$\mathbb{E}X^2 = \mathbb{E}(gg^*)^2 - \frac{1}{(TD)^2} I = \mathbb{E}\|g\|_2^4 \hat{g}\hat{g}^* - \frac{1}{(TD)^2}$$
$$= \frac{1}{(TD)^2}((3d + d(d-1))\frac{1}{d} I - I) = \frac{d+1}{(TD)^2} I$$

Here $\hat{g} := g/\|g\|_2$ and the calculation is done by independence of $\|g\|_2, \hat{g}$. So we also have the variance parameter

$$\sigma^2 = N\|\mathbb{E}X^2\|_{op} = \frac{TD}{d} \frac{d+1}{(TD)^2} \sim \frac{1}{TD}$$

Corollary 8. *We have the following operator norm concentration*

$$\mathbb{P}[\|Q^a - sI_a\|_{op} \geq t] \leq d \exp\left(-\frac{\Omega(t^2 TD)}{1 + t d_a \log D}\right)$$

Since we require $\|\cdot\|_{op}$ error $\ll \frac{1}{d_a \log D}$, if we are happy with $1/poly$ failure probability we require $TD \gg \max_a d_a^2 \log^3 D$.

Remark 2. *Note I'm using $\min_a d_a < \max_a d_a < D$ in a couple places so the* log *term may be slightly sharpened. But the exponent is tight as we require $TD > \max_a d_a^2$ samples for existence/ uniqueness of the solution.*

3.2 Gaussian proof of $\|\mu\|_{op}$

The above method of first thresholding the Gaussians then using Bernstein-style concentration on a bounded random matrix feels a bit square-peg round-hole - y. Turns out there are better results specifically for the case of Gaussian matrices. Recall again that in our setting Q^a is the sum of $N := \frac{nD}{d_a}$ copies of XX^* where $X \sim \mathcal{N}(0, \frac{1}{nD} I_a)$. Note first the following fact which allows us to use these specialized inequalities

Fact 9. *$\sum_{i=1}^N X_i X_i^* \equiv GG^*$ where $G := \{X_1, ..., X_N\}$. This means if we denote $\{\lambda_1, ..., \lambda_d\}$ the spectrum of $\sum_{i=1}^N X_i X_i^*$, this is the same as $\{s_1^2, ..., s_d^2\}$ where $s_j := s_j(G)$ the j-th singular value. By Taylor expansion of $\sqrt{1+x}$ we have:*

$$\lambda_1, \lambda_d(GG^*) \in \frac{1}{d_a}\left(1 \pm \frac{1}{\log d_a}\right) \iff s_1, s_d(G) \in \frac{1}{\sqrt{d_a}}\left(1 \pm \frac{1}{\log d_a}\right)$$

Corollary 10 (Corollary 5.35)**.** *Let $G_{d,N} \in \mathbb{R}^{d\times N}$ for $d < N$ have independent standard Gaussian entries. Then for $t \geq 0$, the following occurs with $\leq 2\exp(-t^2/2)$ failure probability:*

$$\sqrt{N} - \sqrt{d} - t \leq s_d(G) \leq s_1(G) \leq \sqrt{N} + \sqrt{d} + t$$

Corollary 11. *If $nD \gtrsim d_a^2 \log^2 d_a$ then $\|Q^a - \frac{1}{d_a} I_a\|_{op} \ll \frac{1}{d_a \log d_a}$ with failure probability $\leq \exp(-\Omega(d_a))$*

Proof. We have the following with $\leq 2\exp(-t^2/2)$ failure probability:

$$s_1, s_d\left(\frac{1}{\sqrt{nD}} G_{d,N}\right) \in \frac{1}{\sqrt{nD}}\left(\sqrt{\frac{nD}{d_a}} \pm (\sqrt{d_a} + t)\right) = \frac{1}{\sqrt{d_a}}\left(1 \pm \frac{d_a + t\sqrt{d_a}}{\sqrt{nD}}\right)$$

Choosing $t \sim \sqrt{d_a}$ and $nD \gtrsim d_a^2 \log^2 d_a$ gives the required bound. ☐

4 Old robustness

Lemma 12. *If f is λ-strongly convex at I and $\forall a : d_a\|(\nabla f)_a\|_{op} \leq \varepsilon \ll 1/k$, then for Z such that $\forall a : \|Z_a\|_{op} \leq \delta_a \ll 1/k$, the function f at e^Z is $\lambda - O(k\sum_a \delta_a)$ strongly convex.*

Lemma 13. *[CF: what I think this lemma should say] Let $f = f_x$. There is a constant $c > 0$ such that if f is λ-strongly convex at I and that $\|(\nabla f)_a\|_{op} \leq \varepsilon \leq ck^{-1}$ for all $a \in [k]$, then the function f is*

$$\lambda - O(k\sum \|Z_a\|_{op})$$

-strongly convex at $Z \in$ PD provided $\|Z_a\|_{op} \leq ck^{-1}$ for all $a \in [k]$. [CF: just define operator norm on PD?]

The bulk of the work goes towards an intermediate lemma showing that each block $\nabla^2_{ab} f$ of the Hessian changes fairly little on the operator norm ball.

Lemma 14. *For perturbation $v \to \otimes_a e^{\delta_a} \cdot v =: w$ where $\forall a : \|\delta_a\|_{op} \ll 1$, and let $\{\sigma_1^{ab}, \sigma_2^{ab}\}$ be the matrix norm $\|\cdot\|_F \to \|\cdot\|_F$ and matrix norm on subspace $\perp$ to (I, I) for each bipartite part respectively:*

$$\forall a, b : \sigma_2^{ab}(ww^*) - \sigma_2^{ab}(vv^*) \leq O\left(\sum_a \|\delta_a\|_{op}\right) \sigma_1^{ab}(vv^*)$$

The same is true for the diagonal blocks.

[CF: I think we are safe to just say $\|\nabla^2_{ab} f\|_{op} := \|\nabla^2_{ab} f\|_{F \to F}$.]

Lemma 15. *[CF: prev lemma in new notation; not married to the Π's.] Let Π denote the projection to the traceless matrices. There is a constant $c > 0$ such that if $\|Z_a\|_{op} \leq c$ for all $a \in [k]$ we have*

$$\|\nabla^2_{ab} f(Z) \circ \Pi\|_{op} - \|\nabla^2_{ab} f(I) \circ \Pi\|_{op} = O\left(\sum_{a \in [k]} \|Z_a\|_{op} \|\nabla^2_{ab} f\|_{op}\right)$$

for all $a, b \in [k]$.

Proof. [CF: needs to be updated to new notation. At this point in the paper, there's no M.] To lower bound the diagonal block, we just need a spectral lower bound on $\{\rho^a\}$, since $\langle vec(X), \nabla^2_{aa}(vec(X))\rangle := \langle \rho^a, X^2\rangle$.

$$\|e^{\delta_a} \rho^a e^{\delta_a} - Q_a\|_{op} \leq O(\|\delta_a\|_{op}) \|Q_a\|_{op}$$

Now we address a perturbation on $b \neq a$. For a spectral lower bound, we choose test $Z \succeq 0$ and let $\delta := e^{2\delta_b} - I$:

$$\langle e^{\delta_b} \rho e^{\delta_b} - \rho, I_{\overline{a}} \otimes Z_a\rangle = \langle \rho, \delta \otimes Z\rangle = \langle Z, V^* \delta V\rangle$$

Here $V \in \mathbb{R}^{d_b \times d_a}$ is the matricized version of ρ. But now since $Z \succeq 0$, the argument is clear

$$\leq \langle Z, V^* |\delta| V\rangle \leq \|\delta\|_{op} \langle Z, V^* I V\rangle = \|\delta\|_{op} \langle \rho, I_{\overline{a}} \otimes Z\rangle \leq \|\delta\|_{op} \|\rho^a\|_{op} \|Z\|_1$$

The argument for the off-diagonal blocks is similar. We first argue the change is small under perturbations just on those parts.

$$\langle vec(Y), M_v^{ab}(vec(Z))\rangle := \langle vv^*, I_{\overline{ab}} \otimes Z \otimes Y\rangle$$

$$\langle vec(Y), M_w^{ab}(vec(Z))\rangle := \langle ww^*, I_{\overline{ab}} \otimes Z \otimes Y\rangle$$

$$\implies M_w = (e^{\delta_b} \otimes e^{\delta_b}) M_v (e^{\delta_a} \otimes e^{\delta_a})$$

$$\implies \|M_w - M_v\|_{op} \leq O(\|\delta_a\|_{op} + \|\delta_b\|_{op})\|M_v\|_{op}$$

where in the last step we used $\delta \ll 1$. [CF: comment things that we wouldn't want to accidentally leave in, as I have done in the next sentence] The more difficult part of the argument to see [AR: at least for me] is the change caused be some other part $c \neq a, b$. First we define $\delta := e^{2\delta_c} - I$, and test vectors Z, Y:

$$\langle ww^* - vv^*, I_{\overline{ab}} \otimes Z \otimes Y\rangle = \langle vv^*, \delta \otimes Z \otimes Y\rangle = \langle Z \otimes Y, V^* \delta V\rangle$$

Here $V \in \mathbb{R}^{d_c \times d_a d_b}$ is the matricized version of v, i.e. the k-th element of ij-th column is $(V_{ij})_k := v_{ijk}$. Now in order to use our operator norm bounds, we need to deal with cancelations, so we split into positive and negative parts $Z := Z_+ - Z_-, Y := Y_+ - Y_-$:

$$|\langle Z \otimes Y, V^* \delta V\rangle| \leq |\langle Z_\pm \otimes Y_\pm, V^* \delta V\rangle|$$

Now we analyze each of these terms:

$$\leq |\langle Z_\pm \otimes Y_\pm, V^* |\delta| V\rangle| \leq \|\delta\|_{op} |\langle Z_\pm \otimes Y_\pm, V^* V\rangle| = \|\delta\|_{op} |\langle vv^*, I_{\overline{ab}} \otimes Z_\pm \otimes Y_\pm\rangle|$$

Each of these terms we can bound by $\sigma_1^{ab}\|Z\|_F\|Y\|_F$. So by iterating this argument over all c, we get the desired bound. $\square$

Proof of Lemma 13.

$$\langle X, \nabla^2_{aa} X\rangle = \langle \rho^a, X^2\rangle \leq \|\rho^a\|_{op}\|X^2\|_1 = \|\rho^a\|_{op}\|X\|_F^2$$

[AR: I am probably wrong on dimension factors here, but it's the right idea] By the condition on the gradient [TODO: what condition? cref it], we have that

$$\forall a, b : \|\nabla^2_{ab}\|_{op}^2 \leq \|\nabla^2_{aa}\|_{op}\|\nabla^2_{bb}\|_{op} = \|\rho^a\|_{op}\|\rho^b\|_{op} \leq \frac{1+\varepsilon}{d_a d_b}$$

We apply the perturbation lemma to each part sucessively, and if δ are small enough we can assume this bound holds in weaker form $1+\epsilon \leq 2$ for all iterations. The above lemma shows for each part and any test vectors

$$\langle ww^* - vv^*, I_{\overline{ab}} \otimes \frac{Z}{\|Z\|_F} \otimes \frac{Y}{\|Y\|_F}\rangle \leq \frac{O(\sum_a \delta_a)}{\sqrt{d_a d_b}} =: \frac{\delta}{\sqrt{d_a d_b}}$$

Here the suppressed constants are ≤ 7. Therefore the difference between Hessians can be bounded

$$|\langle Y, \nabla^2 f(e^Z) - \nabla^2 f(I), Y\rangle| \leq \delta \left(\sum_a \frac{\|Y_a\|_F^2}{d_a} + \sum_{a \neq b} \frac{\|Y_a\|_F\|Y_b\|_F}{\sqrt{d_a d_b}} \right) \leq k\delta\|Y\|^2$$

$\square$

After one more simple lemma, we will be ready to prove our second strong convexity result, Lemma 13.

Lemma 16 (Lemma 3.6 in [TODO: cite KLR]; [CF: where is this used?]). *[AR: The amount we lose in robustness is related to the worst quadratic form in the whole space (not $\perp I$) since we have to break up into $\pm$ parts.]*

$$\|\nabla^2_{ab}\|^2_{F\to F} \leq \|\nabla^2_{aa}\|_{F\to F}\|\nabla^2_{bb}\|_{F\to F}$$

Proof. [AR: New simple proof:] By convexity we know $\begin{pmatrix} \nabla^2_{aa} & \nabla^2_{ab} \\ \nabla^2_{ba} & \nabla^2_{bb} \end{pmatrix} \succeq 0$. The result follows from e.g. Schur complements. $\square$