[Lectures 10-11]

1 Resilience Beyond Mean Estimation

We have so far focused primarily on mean estimation, first considering information theoretic and then algorithmic issues. We now turn back to information theoretic issues with a focus on generalizing our results from mean estimation to other statistical problems.

Let us recall our general setup: for true (test) distribution p^* and corrupted (train) distribution $\tilde{p}$, we observe samples $X_1, \ldots, X_n$ from $\tilde{p}$ (oblivious contamination, although we can also consider adaptive contamination as in Section **??**). We wish to estimate a parameter θ and do so via en estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$. Our goal is to construct an estimator such that $L(p^*, \hat{\theta})$ is small according to a given loss function L. This was summarized in Figure **??** from Section **??**.

As before, we will start by allowing our estimator $\hat{\theta}$ to directly access the population distribution $\tilde{p}$ rather than samples. Thus we wish to control the error $L(p^*, \hat{\theta}(\tilde{p}))$. Since this is hopeless without further assumptions, we assume that $D(p^*, \tilde{p}) \le \epsilon$ for some distance D, and that p^* lies in some family $\mathcal{G}$.

For now we continue to take $D = \mathsf{TV}$ and focus on more general losses L, corresponding to tasks beyond mean estimation. Two key examples will be:

- **Second moment estimation** in spectral norm, which corresponds to the loss $L(p, S) = \|\mathbb{E}_p[XX^\top] - S\|$.
- **Linear regression**, which corresponds to the loss $L(p, \theta) = \mathbb{E}_{x,y\sim p}[(y - \theta^\top x)^2 - (y - \theta^*(p)^\top x)^2]$. Note that here L measures the *excess predictive loss* so that $L(p, \theta^*(p)) = 0$.

As in the mean estimation case, we will define the modulus of continuity and the family of resilience distributions, and derive sufficient conditions for resilience.

Modulus of continuity. The modulus of continuity generalizes straightforwardly from the mean estimation case. We define

$$\mathfrak{m}(\mathcal{G}, 2\epsilon, L) = \sup_{p,q\in\mathcal{G}:\mathsf{TV}(p,q)\le 2\epsilon} L(p, \theta^*(q)). \tag{1}$$

As before, the modulus $\mathfrak{m}$ upper-bounds the minimax loss. Specifically, consider the projection estimator that outputs $\hat{\theta}(\tilde{p}) = \theta^*(q)$ for any $q \in \mathcal{G}$ with $\mathsf{TV}(\tilde{p}, q) \le \epsilon$. Then the error of $\hat{\theta}$ is at most $\mathfrak{m}$ because $\mathsf{TV}(q, p^*) \le 2\epsilon$ and $p^*, q \in \mathcal{G}$.

Resilience. Generalizing resilience requires more care. Recall that for mean estimation the set of (ρ, ϵ)-resilient distributions was

$$\mathcal{G}^{\mathsf{TV}}_{\mathsf{mean}}(\rho, \epsilon) \stackrel{\text{def}}{=} \big\{p \mid \|\mathbb{E}_r[X] - \mathbb{E}_p[X]\| \le \rho \text{ for all } r \le \frac{p}{1-\epsilon}\big\}. \tag{2}$$

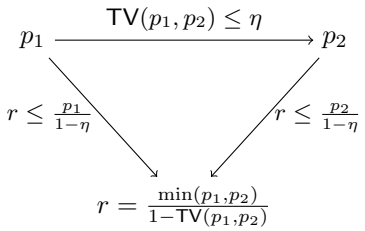
We saw in Section **??** that robust mean estimation is possible for the family $\mathcal{G}_{\mathsf{mean}}$ of resilient distributions; the two key ingredients were the existence of a midpoint distribution and the triangle inequality for $L(p, \theta^*(q)) = \|\mu_p - \mu_q\|$. We now extend the definition of resilience to arbitrary cost functions $L(p, \theta)$ that may not satisfy the triangle inequality. The general definition below imposes two conditions: (1) the parameter $\theta^*(p)$ should do well on all distributions $r \le \frac{p}{1-\epsilon}$, and (2) any parameter that does well on some $r \le \frac{p}{1-\epsilon}$ also does well on p. We measure performance on r with a *bridge function* $B(r, \theta)$, which is often the same as the loss L but need not be.

Definition 1.1 ($\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon)$)**.** Given an arbitrary loss function $L(p, \theta)$, we define $\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon) = \mathcal{G}^{\mathsf{TV}}_{\downarrow}(\rho_1, \epsilon) \cap \mathcal{G}^{\mathsf{TV}}_{\uparrow}(\rho_1, \rho_2, \epsilon)$, where:

$$\mathcal{G}^{\mathsf{TV}}_{\downarrow}(\rho_1, \epsilon) \triangleq \{p \mid \sup_{r\le\frac{p}{1-\epsilon}} B(r, \theta^*(p)) \le \rho_1\}, \tag{3}$$

$$\mathcal{G}^{\mathsf{TV}}_{\uparrow}(\rho_1, \rho_2, \epsilon) \triangleq \{p \mid \text{ for all } \theta, r \le \frac{p}{1-\epsilon}, (B(r, \theta) \le \rho_1 \Rightarrow L(p, \theta) \le \rho_2)\}, \tag{4}$$

The function $B(p, \theta)$ is an arbitrary cost function that serves the purpose of bridging.



$$p_1 \in \mathcal{G}_{\downarrow}^{\mathsf{TV}} \Longrightarrow B(r, \theta^*(p_1)) \le \rho_1 \xRightarrow{p_2 \in \mathcal{G}_{\uparrow}^{\mathsf{TV}}} L(p_2, \theta^*(p_1)) \le \rho_2$$

Figure 1: Midpoint distribution helps bridge the modulus for $\mathcal{G}^{\mathsf{TV}}$.

If we take $B(p, \theta) = L(p, \theta) = \|\mathbb{E}_p[X] - \mathbb{E}_\theta[X]\|, \rho_2 = 2\rho_1$, then this exactly reduces to the resilient set $\mathcal{G}_{\mathsf{mean}}^{\mathsf{TV}}(\rho_1, \epsilon)$ for mean estimation. To see the reduction, note that $\mathcal{G}_{\mathsf{mean}}^{\mathsf{TV}}$ is equivalent to $\mathcal{G}_{\downarrow}^{\mathsf{TV}}$ in Equation (3). Thus we only need to show that $\mathcal{G}_{\uparrow}^{\mathsf{TV}}$ is a subset of $\mathcal{G}_{\downarrow}^{\mathsf{TV}}$. By our choice of B, L and ρ_2, the implication condition in $\mathcal{G}_{\uparrow}^{\mathsf{TV}}$ follows from the triangle inequality..

We will show that $\mathcal{G}^{\mathsf{TV}}$ is *not too big* by bounding its modulus of continuity, and that it is *not too small* by exhibiting reasonable sufficient conditions for resilience.

Not too big: bounding $\mathfrak{m}$. We show that the designed $\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon)$ has small modulus of continuity (and thus population minimax limit) in the following theorem:

Theorem 1.2. *For $\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon)$ in Definition 1.1, we have $\mathfrak{m}(\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon), \epsilon) \le \rho_2$.*

Proof. As illustrated in Figure 1, we still rely on the midpoint distribution r to bridge the modulus. Consider any p_1, p_2 satisfying $\mathsf{TV}(p_1, p_2) \le \epsilon$. Then there is a midpoint r such that $r \le \frac{p_1}{1-\epsilon}$ and $r \le \frac{p_2}{1-\epsilon}$. From the fact that $p_1 \in \mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon) \subset \mathcal{G}_{\downarrow}^{\mathsf{TV}}(\rho_1, \epsilon)$, we have $B(r, \theta^*(p_1)) \le \rho_1$. From this and the fact that $p_2 \in \mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon) \subset \mathcal{G}_{\uparrow}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon)$, we then have $L(p_2, \theta^*(p_1)) \le \rho_2$. Since p_1 and p_2 are arbitrary, this bounds the modulus of continuity by ρ_2. □

Not too small: concrete examples. We next show that $\mathcal{G}^{\mathsf{TV}}$ yields sensible conditions for second moment estimation and linear regression. We start with second moment estimation:

Proposition 1.3. *Let $B(p, S) = L(p, S) = \|\mathbb{E}_p[XX^\top] - S\|$, and let p be a distribution on $\mathbb{R}^d$ such that $p \in \mathcal{G}_{\mathsf{mom},k})(\sigma)$, i.e. p^* has bounded kth moments. Then assuming $k > 2$, we have $p \in \mathcal{G}^{\mathsf{TV}}(\rho, 2\rho, \epsilon)$ for $\rho = \mathcal{O}(\sigma^2 \epsilon^{1-2/k})$.*

This is essentially the same statement as for mean estimation, except with $\sigma^2\epsilon^{1-2/k}$ instead of $\sigma\epsilon^{1-1/k}$.

Proof. First we show that $p \in \mathcal{G}^{\downarrow}(\rho, \epsilon)$, for which we need to show that

$$\|\mathbb{E}_r[XX^\top] - \mathbb{E}_p[XX^\top]\| \le \rho \text{ for all } r \le \frac{p}{1-\epsilon}. \tag{5}$$

Letting $Y = XX^\top$, this asks that Y is resilient in operator norm, which in turn asks that $\langle Y, Z\rangle$ is resilient for any $\|Z\|_* \le 1$, where $\|\cdot\|_*$ is dual to the operator norm. Recalling that the operator norm is the maximum singular value, it turns out that $\|\cdot\|_*$ is the *nuclear norm*, or the sum of the singular values. Thus for $Z = U\Lambda V^\top$ we have $\|Z\|_* = \sum_i \Lambda_{ii}$. (Proving this duality requires some non-trivial but very useful matrix inequalities that we provide at the end of this section.)

Conveniently, the extreme points of the nuclear norm ball are exactly rank-one matrices of the form $\pm vv^\top$ where $\|v\|_2 = 1$. Thus we exactly need that $\langle v, X\rangle^2$ is resilience for all v. Fortunately we have that $\mathbb{E}[|\langle v, X\rangle^2 - \mathbb{E}[\langle v, X\rangle^2]|^{k/2}] \le \mathbb{E}[|\langle v, X\rangle|^k] \le \sigma^k$, so p is (ρ_1, ϵ)-resilient with $\rho_1 = \sigma^2\epsilon^{1-2/k}$, which gives that $p \in \mathcal{G}^{\downarrow}$.

Next we need to show that $p \in \mathcal{G}^{\uparrow}$. We want

$$\|\mathbb{E}_r[XX^\top] - S\| \le \rho_1 \implies \|\mathbb{E}_p[XX^\top] - S\| \le \rho_2 \text{ whenever } r \le \frac{p}{1-\epsilon}, \tag{6}$$

but this is the same as $\rho_2 - \rho_1 \le \|\mathbb{E}_r[XX^\top] - \mathbb{E}_p[XX^\top]\|$, and we already know that the right-hand-side is bounded by ρ_1, so we can take $\rho_2 = 2\rho_1$, which proves the claimed result. □

We move on to linear regression. In the proof for second moment estimation, we saw that $p \in \mathcal{G}^{\uparrow}$ was essentially implied by $p \in \mathcal{G}^{\downarrow}$. This was due to the symmetry of the second moment loss together with the triangle inequality for $\|\cdot\|$, two properties that we don't have in general. The proof for second moment estimation will require somewhat more different proofs for $\mathcal{G}^{\uparrow}$ and $\mathcal{G}^{\downarrow}$. For simplicity we state the result only for fourth moments:

Proposition 1.4. *For a distribution p on $\mathbb{R}^d \times \mathbb{R}$, let $B(p,\theta) = L(p,\theta) = \mathbb{E}_p[(y - \langle\theta, x\rangle)^2 - (y - \langle\theta^*(p), x\rangle)^2]$. Let $Z = Y - \langle\theta^*(p), X\rangle$ and suppose that the following two conditions holds:*

$$\mathbb{E}_p[XZ^2X^\top] \preceq \sigma^2\mathbb{E}[XX^\top], \tag{7}$$

$$\mathbb{E}_p[\langle X, v\rangle^4] \le \kappa\mathbb{E}_p[\langle X, v\rangle^2]^2 \text{ for all } v. \tag{8}$$

Then $p \in \mathcal{G}^{\mathsf{TV}}(\rho, 5\rho, \epsilon)$ for $\rho = 2\sigma^2\epsilon$ as long as $\epsilon(\kappa - 1) \le \frac{1}{6}$ and $\epsilon \le \frac{1}{8}$.

Let us interpret the two conditions. First, as long as X and Z are independent (covariates are independent of noise), we have $\mathbb{E}_p[XZ^2X^\top] = \mathbb{E}[Z^2]\mathbb{E}[XX^\top]$, so in that case σ^2 is exactly a bound on the noise Z. Even when X and Z are not independent, the first condition holds when Z has bounded 4th moment.

The second condition is a *hypercontractivity condition* stating that the fourth moments of X should not be too large compared to the second moments. It is a bit unusual from the perspective of mean estimation, because it does not require X to be well-concentrated, but only well-concentrated relative to its variance. For regression, this condition makes sense because κ bounds how close the covariates are to being rank-deficient (the worst-case is roughly an ϵ-mass at some arbitrary distance $t/\sqrt{\epsilon}$, which would have second moment t^2 and fourth moment t^4/ϵ, so we roughly want $\kappa < 1/\epsilon$). We will show later that such a hypercontractivity condition is needed, i.e. simply assuming sub-Gaussianity (without making it relative to the variance) allows for distributions that are hard to robustly estimate due to the rank-deficiency issue.

Proof. First note that $L(p,\theta) = (\theta - \theta^*(p))^\top S_p(\theta - \theta^*(p))$, where $S_p = \mathbb{E}_p[XX^\top]$, and analogously for $L(r,\theta)$. At a high level our strategy will be to show that $\theta^*(r) \approx \theta^*(p)$ and $S_r \approx S_p$, and then use this to establish membership in $\mathcal{G}^{\downarrow}$ and $\mathcal{G}^{\uparrow}$.

We first use the hypercontractivity condition to show that $S_r \approx S_p$. We have

$$\mathbb{E}_r[\langle v, X\rangle^2] \ge \mathbb{E}_p[\langle v, X\rangle^2] - \frac{1}{1-\epsilon}\sqrt{\epsilon\mathsf{Var}_p[\langle v, X\rangle^2]} \tag{9}$$

$$= \mathbb{E}_p[\langle v, X\rangle^2] - \frac{1}{1-\epsilon}\sqrt{\epsilon(\mathbb{E}_p[\langle v, X\rangle^4] - \mathbb{E}_p[\langle v, X\rangle^2]^2)} \tag{10}$$

$$\ge \mathbb{E}_p[\langle v, X\rangle^2] - \frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)}\mathbb{E}_p[\langle v, X\rangle^2] \tag{11}$$

$$= (1 - \frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)})\mathbb{E}_p[\langle v, X\rangle^2]. \tag{12}$$

Thus $S_r \succeq (1 - \frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)})S_p$, and similarly $S_r \preceq (1 + \frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)})S_p$. Assuming $\epsilon \le \frac{1}{8}$ and $\epsilon(\kappa-1) \le \frac{1}{6}$, we have $\frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)} \le \frac{8}{7}\sqrt{1/6} < \frac{1}{2}$, and so $\frac{1}{2}S_p \preceq S_r \preceq \frac{3}{2}S_p$.

We now turn to $\mathcal{G}^{\uparrow}$ and $\mathcal{G}^{\downarrow}$. A useful relation is $\theta^*(p) = S_p^{-1}\mathbb{E}_p[XY]$, and $\theta^*(r) - \theta^*(p) = S_r^{-1}\mathbb{E}_r[XZ]$. To prove that $p \in \mathcal{G}^{\downarrow}$ we need to show that $(\theta^*(r) - \theta^*(p))^\top S_r(\theta^*(r) - \theta^*(p))$ is small. We have

$$(\theta^*(r) - \theta^*(p))^\top S_r(\theta^*(r) - \theta^*(p)) \le \frac{3}{2}(\theta^*(r) - \theta^*(p))^\top S_p(\theta^*(r) - \theta^*(p)) \tag{13}$$

$$= \frac{3}{2}\mathbb{E}_r[XZ]^\top S_p^{-1}\mathbb{E}_r[XZ] = \frac{3}{2}\|\mathbb{E}_r[S_p^{-1/2}XZ] - \mathbb{E}_p[S_p^{-1/2}XZ]\|_2^2. \tag{14}$$

This final condition calls for $S_p^{-1/2} XZ$ to be resilient, and bounded variance of this distribution can be seen to exactly correspond to the condition $\mathbb{E}[XZ^2X^\top] \preceq \sigma^2 \mathbb{E}[XX^\top]$. Thus we have resilience with $\rho = \frac{3\sigma^2\epsilon}{2(1-\epsilon)^2} \le 2\sigma^2\epsilon$ (since $\epsilon < \frac{1}{8}$).

Now we turn to $\mathcal{G}^\uparrow$. We want that $(\theta - \theta^*(r))^\top S_r(\theta - \theta^*(r)) \le \rho$ implies $(\theta - \theta^*(p))^\top S_p(\theta - \theta^*(p)) \le 5\rho$. By the triangle inequality we have

$$\sqrt{(\theta - \theta^*(p))^\top S_p(\theta - \theta^*(p))} \le \sqrt{(\theta - \theta^*(r))^\top S_p(\theta - \theta^*(r))} + \sqrt{(\theta^*(r) - \theta^*(p))^\top S_p(\theta^*(r) - \theta^*(p))} \quad (15)$$

$$\le \sqrt{2(\theta - \theta^*(r))^\top S_r(\theta - \theta^*(r))} + \sqrt{(4/3)\sigma^2\epsilon} \quad (16)$$

$$\le \sqrt{2\rho} + \sqrt{(4/3)\sigma^2\epsilon} = \sqrt{\rho}(\sqrt{2} + \sqrt{2/3}) \le \sqrt{5\rho}, \quad (17)$$

which completes the proof. □

Lower bound. TBD

Proving that nuclear norm is dual to operator norm. Here we establish a series of matrix inequalities that are useful more broadly, and use these to analyze the nuclear norm. The first allows us to reduce dot products of arbitrary matrices to symmetric PSD matrices:

Proposition 1.5. *For any (rectangular) matrices A, B of equal dimensions, we have*

$$\langle A, B\rangle^2 \le \langle (AA^\top)^{1/2}, (B^\top B)^{1/2}\rangle \langle (AA^\top)^{1/2}, (BB^\top)^{1/2}\rangle. \quad (18)$$

In a sense, this is like a "matrix Cauchy-Schwarz".

Proof. We first observe that $\begin{bmatrix} (AA^\top)^{1/2} & A \\ A^\top & (A^\top A)^{1/2} \end{bmatrix} \succeq 0$. This is because, if $A = U\Lambda V^\top$ is the singular value decomposition, we can write the above matrix as $\begin{bmatrix} U\Lambda U^\top & U\Lambda V^\top \\ V\Lambda U^\top & V\Lambda V^\top \end{bmatrix}$, which is PSD because it can be factorized as $[U; V]\Lambda[U; V]^\top$. More generally this is true if we multiply $(AA^\top)^{1/2}$ by λ and $(A^\top A)^{1/2}$ by $\frac{1}{\lambda}$. We therefore have

$$\left\langle \begin{bmatrix} \lambda(AA^\top)^{1/2} & A \\ A^\top & \frac{1}{\lambda}(A^\top A)^{1/2} \end{bmatrix}, \begin{bmatrix} \lambda(BB^\top)^{1/2} & -B \\ -B^\top & \frac{1}{\lambda}(B^\top B)^{1/2} \end{bmatrix} \right\rangle \ge 0, \quad (19)$$

since both terms in the inner product are PSD. This gives $\lambda^2\langle (AA^\top)^{1/2}, (BB^\top)^{1/2}\rangle + \frac{1}{\lambda^2}\langle (A^\top A)^{1/2}, (B^\top B)^{1/2}\rangle \ge 2\langle A, B\rangle$. Optimizing λ yields the claimed result. □

Next we show:

Theorem 1.6. *If A and B are matrices of the same dimensions with (sorted) lists of singular values $\sigma_1, \dots, \sigma_n$ and $\tau_1, \dots, \tau_n$, then*

$$\langle A, B\rangle \le \sum_{i=1}^n \sigma_i \tau_i. \quad (20)$$

This says that the dot product between two matrices is bounded by the dot product between their sorted singular values.

Proof. By Proposition 1.5, it suffices to show this in the case that A and B are both PSD and σ, τ are the eigenvalues. Actually we will only need A and B to be symmetric (which implies that, oddly, the inequality can hold even if some of the σ_i and τ_i are negative).

By taking similarity transforms we can assume without loss of generality that $A = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$. We thus wish to prove that $\sum_{i=1}^n \sigma_i B_{ii} \le \sum_{i=1}^n \sigma_i \tau_i$, where τ_i are the eigenvalues of B. We make use of the following lemma:

Lemma 1.7. *For all $1 \le k \le n$, we have $\sum_{i=1}^k B_{ii} \le \sum_{i=1}^k \tau_i$.*

Proof. Let B_k be the $k \times k$ top-left submatrix of B. Then $\sum_{i=1}^{k} B_{ii} = \mathrm{tr}(B_k)$ is the sum of the eigenvalues of B_k. We will show that the jth largest eigenvalue of B_k is smaller than the jth largest eigenvalue of B (this is a special case of the *Cauchy interlacing theorem*). We prove this using the min-max formulation of eigenvalues: $\lambda_i(M) = \min_{W:\dim(W)=i-1} \max_{v \in W^\perp, \|v\|_2 \le 1} v^\top M v$. Let W^* be the W that attains the min for $\lambda_j(B)$, and let P_k denote projection onto the first k coordinates. We have

$$\begin{aligned}
\lambda_j(B_k) &= \min_{W:\dim(W)=i-1} \max_{v \in W^\perp : \|v\|_2 \le 1} v^\top B_k v && (21)\\
&\le \max_{v \in (W^*)^\perp : \|v\|_2 \le 1} (P_k v)^\top B_k (P_k v) && (22)\\
&\le \max_{v \in (W^*)^\perp : \|v\|_2 \le 1} v^\top B v = \lambda_j(B), && (23)
\end{aligned}$$

which proves the lemma. □

Now with the lemma in hand we observe that, if we let $\sigma_{n+1} = 0$ for convenience, we have

$$\begin{aligned}
\sum_{i=1}^{n} \sigma_i B_{ii} &= \sum_{i=1}^{n} (\sigma_i - \sigma_{i+1})(B_{11} + \cdots + B_{ii}) && (24)\\
&\le \sum_{i=1}^{n} (\sigma_i - \sigma_{i+1})(\tau_1 + \cdots + \tau_i) && (25)\\
&= \sum_{i=1}^{n} \sigma_i \tau_i, && (26)
\end{aligned}$$

which yields the desired result. In the above algebra we have used *Abel summation*, which is the discrete version of integration by parts. □

Now that we have Theorem 1.6 in hand, we can easily analyze the operator and nuclear norms. Letting $\vec{\sigma}(A)$ denote the vector of non-decreasing singular values of A, we have

$$\langle Y, Z \rangle \le \langle \vec{\sigma}(Y), \vec{\sigma}(Z) \rangle \le \|\vec{\sigma}(Y)\|_\infty \|\vec{\sigma}(Z)\|_1. \tag{27}$$

This shows that the dual of the operator norm is at most the nuclear norm, since $\|\vec{\sigma}(Z)\|_1$ is the nuclear norm of Z. But we can achieve equality when $Y = U \Lambda V^\top$ by taking $Z = u_1 v_1^\top$ (then $\|Z\|_* = 1$ while $\langle Y, Z \rangle = \Lambda_{11} = \|Y\|$). So operator and nuclear norm are indeed dual to each other.