

A Properties of Huber distribution

A.1 Normalizing factor of Huber distribution

Recall

$$p(\mathbf{x} \mid \mathbf{v}, \Lambda, \delta) \propto \exp \left(-h_\delta \left(\|\Lambda^{1/2}(\mathbf{x} - \mathbf{v})\|_2 \right) \right) \quad (14)$$

To make this integrate to 1 we need a normalizing factor Z . It will be a function of the parameters of the distribution, i.e. \mathbf{v} , Λ and δ . First we consider how \mathbf{v} and Λ influence the normalizing factor. By doing the variable substitution $\mathbf{y} = \Lambda^{1/2}(\mathbf{x} - \mathbf{v})$ we get the following

$$1 = \frac{1}{Z(\Lambda, \mathbf{v}, \delta)} \int_{\mathbf{x} \in \mathbb{R}^d} p(\mathbf{x} \mid \mu, \Lambda, \delta) d\mathbf{x} \quad (15)$$

$$= \frac{1}{Z(\Lambda, \mathbf{v}, \delta) |\Lambda^{1/2}|} \int_{\mathbf{y} \in \mathbb{R}^d} p(\mathbf{y} \mid 0, I, \delta) d\mathbf{y} \quad (16)$$

$$= \frac{c_d(\delta)}{Z(\Lambda, \mathbf{v}, \delta) |\Lambda^{1/2}|} \quad (17)$$

Solving this with respect to Z gives

$$Z(\Lambda, \mathbf{v}, \delta) = \frac{c_d(\delta)}{|\Lambda^{1/2}|} \quad (18)$$

By using the normalizing factor we get the pdf for the distribution

$$p(\mathbf{x} \mid \mathbf{v}, \Lambda, \delta) = \frac{|\Lambda|^{1/2}}{c_d(\delta)} \exp \left(-h_\delta \left(\|\Lambda^{1/2}(\mathbf{x} - \mathbf{v})\|_2 \right) \right) \quad (19)$$

We find the expression of $c_d(\delta)$ by evaluating the integral which defines it. By doing a change to spherical coordinates and using radial symmetry we get

$$c_d(\delta) = \int_0^\infty |S_{d-1}| r^{d-1} \exp(-h_\delta(r)) dr \quad (20)$$

$$= |S_{d-1}| \left(\int_0^\delta r^{d-1} \exp(-r^2/2) dr + \int_\delta^\infty r^{d-1} \exp(-\delta r + \delta^2/2) dr \right) \quad (21)$$

$$= |S_{d-1}| (a(d-1, \delta) + \exp(\delta^2/2) b(d-1, \delta)) \quad (22)$$

Where $|S_d|$ is the volume of a d dimensional unit sphere $S_d = \{x \in \mathbb{R}^{d+1} : \|x\|_2 = 1\}$. a and b are defined by the two integrals.

We first notice

$$a(0, \delta) = \int_0^\delta \exp(-r^2/2) dr = \sqrt{\frac{\pi}{2}} \operatorname{erf}(\delta/\sqrt{2}) \quad (23)$$

and

$$a(1, \delta) = \int_0^{\delta} r * \exp(-r^2/2) dr = [-\exp(-r^2/2)]_0^{\delta} = 1 - \exp(\delta^2/2) \quad (24)$$

by performing integration by parts we get

$$a(n, \delta) = \int_0^{\delta} r^{n-1} r * \exp(-r^2/2) dr \quad (25)$$

$$= [-r^{n-1} \exp(-r^2/2)]_0^{\delta} + \int_0^{\delta} (n-1) r^{n-2} \exp(-r^2/2) \quad (26)$$

$$= -\delta^{n-1} \exp(-\delta^2/2) + (n-1) a(n-2, \delta) \quad (27)$$

from this we have recursively defined $a(n, \delta)$ for all values.

similarly for b

$$b(0, \delta) = \int_{\delta}^{\infty} \exp(-\delta r) dr = \frac{\exp(-\delta^2)}{\delta} \quad (28)$$

$$b(n, \delta) = \int_{\delta}^{\infty} r^n \exp(-\delta r) dr \quad (29)$$

$$= \left[-\frac{r^n}{\delta} \exp(-\delta r) \right]_{\delta}^{\infty} + \int_{\delta}^{\infty} n r^{n-1} \exp(-\delta r) dr \quad (30)$$

$$= \delta^{n-1} \exp(-\delta^2) + n b(n-1, \delta) \quad (31)$$

we have now defined a and b for all n and δ , thereby also the normalizing the normalizing constant for all d and δ .

A.2 Variance of huber distribution

The variance for $p(\mathbf{x} | 0, I, \delta)$ can be found by

$$E(\|\mathbf{X}\|_2^2) = \frac{|S_{d-1}|}{c_d(\delta)} \int_0^{\infty} r^{d+1} \exp(-h_{\delta}(r)) dr = \frac{a(d+1, \delta) + \exp(\delta^2/2) b(d+1, \delta)}{a(d-1, \delta) + \exp(\delta^2/2) b(d-1, \delta)} \quad (32)$$

Due to symmetry we know that $\text{Var}(\mathbf{X})$ is a diagonal matrix and $\text{tr}(\text{Var}(\mathbf{X})) = E(\|\mathbf{X}\|_2^2)$ therefore

$$\text{Var}(\mathbf{X}) = \frac{a(d+1, \delta) + \exp(\delta^2/2) b(d+1, \delta)}{d(a(d-1, \delta) + \exp(\delta^2/2) b(d-1, \delta))} \quad (33)$$

We get the following after doing a variable substitution $\mathbf{y} = \Sigma^{1/2}\mathbf{x}$

$$\text{Var}(\mathbf{Y}) = \frac{a(d+1, \delta) + \exp(\delta^2/2)b(d+1, \delta)}{d(a(d-1, \delta) + \exp(\delta^2/2)b(d-1, \delta))} \Sigma \quad (34)$$

The expected distance between the mean and a sample will then be

$$E(\|\mathbf{Y}\|_2^2) = E(\text{tr}(\mathbf{Y}^T \mathbf{Y})) = E(\text{tr}(\mathbf{Y} \mathbf{Y}^T)) = \quad (35)$$

$$\frac{a(d+1, \delta) + \exp(\delta^2/2)b(d+1, \delta)}{d(a(d-1, \delta) + \exp(\delta^2/2)b(d-1, \delta))} \text{tr}(\Sigma) \quad (36)$$

Specifically for $d = 2, \delta = 1$

$$E(\|\mathbf{Y}\|_2^2) \approx 3.07 \text{tr}(\Sigma) \quad (37)$$

B Equation for gradients when applying function on eigenvalues

If we have the square symmetric matrix B with eigendecomposition $B = V^T D V$ and define $A = G(B) = V^T \text{diag}(g(D_{1,1}), g(D_{2,2}), \dots, g(D_{d,d})) V$ where g is a differentiable function $g: \mathbb{R} \rightarrow \mathbb{R}$. Then the gradient of a function L can be computed with respect to B through the following equation.

$$\frac{\partial L}{\partial B} = V^T (V (\frac{\partial L}{\partial A}) V^T \circ K(D, g)) V \quad (38)$$

where

$$K(D, g)_{i,j} = \begin{cases} \frac{g(D)_{i,i} - g(D)_{j,j}}{D_{i,i} - D_{j,j}} & \text{if } D_{i,i} \neq D_{j,j} \\ g'(D_{i,i}) & \text{otherwise} \end{cases} \quad (39)$$

and \circ is a elementwise multiplication. This expression is similar to the expressions in [46], except it handles the case when different eigenvalues are equal as well.

Note that $\frac{\partial L}{\partial A}$ and $\frac{\partial L}{\partial B}$ needs to be symmetric matrices since A and B are symmetric.

B.1 Proof

B.1.1 Reduce proof to diagonal matrices

given a matrix $B = V^T D V$ pick the constant $\hat{V} = V$, note V is a variable dependent on B while \hat{V} is constant.

Define $C = \hat{V} A \hat{V}^T \implies \hat{V}^T C \hat{V} = A$ and $E = \hat{V} B \hat{V}^T \implies \hat{V}^T E \hat{V} = B$

First

$$\frac{\partial L}{\partial C_{i,j}} = \sum_{m=1}^d \sum_{n=1}^d \hat{V}_{i,m} \frac{\partial L}{\partial A_{m,n}} (\hat{V}^T)_{n,j} = (\hat{V} \frac{\partial L}{\partial A} \hat{V}^T)_{i,j} \quad (40)$$

The same holds for any multiplication of constant matrices.

Such as

$$\frac{\partial L}{\partial B_{i,j}} = (\hat{V}^T \frac{\partial L}{\partial E} \hat{V})_{i,j} \quad (41)$$

Since C and E are diagonal this further simplifies our proof.

B.1.2 Differentiation of diagonal elements

We will use the single entry matrix $J^{i,j}$ in following sections. The dimension of this matrix is implicit based on context.

$$J_{m,n}^{i,j} = \mathbb{1}(i = m \wedge j = n) \quad (42)$$

where $\mathbb{1}$ is the indicator function.

If E is diagonal then $F = E + J^{i,i}\epsilon$ is trivially diagonal as well, therefore

$$\lim_{\epsilon \rightarrow 0} \frac{G(E) - G(F)}{\epsilon} = J^{i,i} g'(E_{i,i}) \forall i \quad (43)$$

Since $C = G(E)$ we get

$$\frac{\partial C}{\partial E_{i,i}} = J^{i,i} g'(E_{i,i}) \quad (44)$$

B.1.3 Differentiation of non-diagonal elements

Let's consider how $g(E)$ changes when we change the element of row i and column j. Since E is diagonal this will only affect the i:th and j:th eigenvalues and eigenvectors. Without loss of generality we can analyze the case when we change the non-diagonal elements of a 2×2 matrix.

$$E = \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \quad (45)$$

First we analyze the case when $x \neq y$

We can find the eigenvalues of $E + \epsilon(J^{1,2} + J^{2,1})$ by solving $|E + \epsilon(J^{1,2} + J^{2,1}) - \lambda I| = 0$ for λ

The solution of this is

$$\lambda = \frac{x+y}{2} \pm \sqrt{\left(\frac{x-y}{2}\right)^2 + \epsilon^2} = \frac{x+y}{2} \pm \left(\frac{|x-y|}{2} + \frac{\epsilon^2}{|x-y|} + \mathcal{O}(\epsilon^3)\right) = \quad (46)$$

The first step can be done by completing the square and the second step is the first terms of the maclaurin series.

Assume $x > y$ then solve for eigenvectors to get

$$\mathbf{v}^T \left(x - \frac{x+y}{2} - \frac{|x-y|}{2} - \frac{\epsilon^2}{|x-y|} + \mathcal{O}(\epsilon^3) \right) \epsilon = \mathbf{v}^T \left(-\frac{\epsilon^2}{|x-y|} + \mathcal{O}(\epsilon^3) \right) \epsilon \quad (47)$$

Solving for \mathbf{v} we get

$$\mathbf{v} = \left[1, \frac{\epsilon}{|x-y|} + \mathcal{O}(\epsilon^2) \right] \quad (48)$$

The normalizing factor for \mathbf{v} will be $1 + \mathcal{O}(\varepsilon^2)$ so it will not influence the limit of the derivative. If $y > x$ the sign of the epsilon term would change.

Our new basis is now

$$\frac{\begin{bmatrix} 1 & -\frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) \\ \frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) & 1 \end{bmatrix}}{(1 + \mathcal{O}(\varepsilon^2))^2} \quad (49)$$

Putting it together

$$E + \varepsilon(J^{1,2} + J^{2,1}) \quad (50)$$

$$= \frac{1}{(1 + \mathcal{O}(\varepsilon^2))^2} \begin{bmatrix} 1 & -\frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) \\ \frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) & 1 \end{bmatrix} \quad (51)$$

$$\times \begin{bmatrix} x + \mathcal{O}(\varepsilon^2) & 0 \\ 0 & y + \mathcal{O}(\varepsilon^2) \end{bmatrix} \quad (52)$$

$$\times \begin{bmatrix} 1 & \frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) \\ -\frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) & 1 \end{bmatrix} \quad (53)$$

Where \times is the standard matrix multiplication. Applying g on the diagonal terms gives

$$G(E + \varepsilon(J^{1,2} + J^{2,1})) \quad (54)$$

$$= \frac{1}{(1 + \mathcal{O}(\varepsilon^2))^2} \begin{bmatrix} 1 & -\frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) \\ \frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) & 1 \end{bmatrix} \quad (55)$$

$$\times \begin{bmatrix} g(x) + \mathcal{O}(\varepsilon^2) & 0 \\ 0 & g(y) + \mathcal{O}(\varepsilon^2) \end{bmatrix} \quad (56)$$

$$\times \begin{bmatrix} 1 & \frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) \\ -\frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) & 1 \end{bmatrix} \quad (57)$$

$$= \frac{1}{(1 + \mathcal{O}(\varepsilon^2))^2} \begin{bmatrix} 1 & -\frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) \\ \frac{\varepsilon}{x-y} + \mathcal{O}(\varepsilon^2) & 1 \end{bmatrix} \quad (58)$$

$$\times \begin{bmatrix} g(x) + \mathcal{O}(\varepsilon^2) & \varepsilon \frac{g(x)}{x-y} + \mathcal{O}(\varepsilon^3) \\ -\varepsilon \frac{g(y)}{x-y} + \mathcal{O}(\varepsilon^3) & g(y) + \mathcal{O}(\varepsilon^2) \end{bmatrix} \quad (59)$$

$$= \frac{1}{(1 + \mathcal{O}(\varepsilon^2))^2} \begin{bmatrix} g(x) + \mathcal{O}(\varepsilon^2) & \varepsilon \frac{g(x) - g(y)}{x-y} + \mathcal{O}(\varepsilon^3) \\ \varepsilon \frac{g(x) - g(y)}{x-y} + \mathcal{O}(\varepsilon^3) & g(y) + \mathcal{O}(\varepsilon^2) \end{bmatrix} \quad (60)$$

The first step comes from the fact that g is continuous. The other two steps are matrix multiplications.

From this it is obvious that

$$\frac{\partial G(E)}{\partial E_{1,2}} = \frac{g(x) - g(y)}{x-y} (J^{1,2} + J^{2,1}) \quad (61)$$

Note $E_{1,2} = E_{2,1}$ since E is symmetric.

Differentiation of non-diagonal when $x=y$ We do the same procedure and solve the eigenvalues to be

$$\lambda = x \pm \varepsilon \quad (62)$$

We solve for eigenvectors and get

$$\mathbf{v}^T [x - x - \varepsilon, \varepsilon] = 0 \quad (63)$$

which gives

$$\mathbf{v} = \begin{bmatrix} \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \end{bmatrix} \quad (64)$$

Therefore

$$\begin{bmatrix} x & \varepsilon \\ \varepsilon & x \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} x + \varepsilon & 0 \\ 0 & x - \varepsilon \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \quad (65)$$

This is also trivially verified by matrix multiplication.

$$G\begin{pmatrix} x & \varepsilon \\ \varepsilon & x \end{pmatrix} \quad (66)$$

$$= \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \quad (67)$$

$$\times \begin{bmatrix} g(x) + \varepsilon g'(x) + \mathcal{O}(\varepsilon^2) & 0 \\ 0 & g(x) - \varepsilon g'(x) + \mathcal{O}(\varepsilon^2) \end{bmatrix} \quad (68)$$

$$\times \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \quad (69)$$

$$= G\begin{pmatrix} g(x) & \varepsilon g'(x) \\ \varepsilon g'(x) & g(x) \end{pmatrix} + \mathcal{O}(\varepsilon^2) \quad (70)$$

From this we see that

$$\frac{\partial g(E)}{\partial E_{1,2}} = g'(x)(J^{1,2} + J^{2,1}) \quad (71)$$

When $x = y$

B.1.4 Wrapping up the proof

From the earlier argument this will now hold for all square diagonal matrices

By combining equations 71, 61 and 44 we get

$$\frac{\partial L}{\partial E_{i,j}} = \left(\frac{\partial L}{\partial C_{i,j}} \right) * \begin{cases} \frac{g(E_{i,i}) - g(E_{j,j})}{E_{i,i} - E_{j,j}} & \text{if } E_{i,i} \neq E_{j,j} \\ g'(E_{i,i}) & \text{otherwise} \end{cases} \quad (72)$$

By combining equations 40, 41 and 72 we can construct a proof for equation 38 for every matrix B.

B.1.5 Final comments

In practice we use the gradient when the two eigenvalues are sufficiently close instead of identical to avoid numerical instability.

C Loss

In this section we prove that under the assumption that $\|\mathbf{x}\|_2$ is bounded, our suggested loss has bounded gradients, is convex for the convex set when all eigenvalues of A are larger than θ . and has bounded Hessians for the same set.

From now on we will only analyze the case $\delta = 1$ since that is the value we use for all experiments.

Recall that our loss is parameterized as

$$p(\mathbf{x} \mid \mathbf{v}, A) = \frac{|A|}{c_d(1)} \exp(-h_1(\|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2)) \quad (73)$$

The negative log likelihood of this function, denoted \mathfrak{L} is:

$$\mathfrak{L}(\mathbf{x}, \mathbf{v}, A) = -\log(|A|) + \log(c_d(1)) + h(\|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2) \quad (74)$$

What remains is to show that $-\log(|A|)$ and $h(\|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2)$ have these properties with respect to A and \mathbf{v} . Note this is stronger than convex with respect to the two variables individually. Since we need

$$\mathfrak{L}(\mathbf{x}, \lambda \mathbf{v}_1 + (1 - \lambda) \mathbf{v}_2, \lambda A_1 + (1 - \lambda) A_2) \leq \lambda \mathfrak{L}(\mathbf{x}, \mathbf{v}_1, A_1) + (1 - \lambda) \mathfrak{L}(\mathbf{x}, \mathbf{v}_2, A_2) \quad (75)$$

C.1 Study of diagonal remapping function

This section is for future reference in the proof. Recall that the function we apply on eigenvalues is

$$g(\lambda) = \begin{cases} \lambda & \text{if } \lambda > \theta \\ \theta \exp(\lambda/\theta - 1) & \text{otherwise} \end{cases} \quad (76)$$

$$g'(\lambda) = \begin{cases} 1 & \text{if } \lambda > \theta \\ \exp(\lambda/\theta - 1) & \text{otherwise} \end{cases} \quad (77)$$

$$g''(\lambda) = \begin{cases} 0 & \text{if } \lambda > \theta \\ \frac{1}{\theta} \exp(\lambda/\theta - 1) & \text{otherwise} \end{cases} \quad (78)$$

g is continuous, has continuous gradients and is convex since the second derivative is positive almost everywhere and the gradient is continuous where the second derivative is undefined.

The derivative of g is always between 0 and 1. For this reason

$$0 \leq (g(x) - g(y))/(x - y) \leq 1 \quad (79)$$

For this reason when backpropagating through this function the gradient magnitude w.r.t. Frobenius norm is guaranteed to decrease, since we do a componentwise multiplication with values between 0 and 1. Therefore if the gradient with respect to A is bounded then the gradient with respect to B will be bounded too. since the mapping from network output to B preserves norms this means that the gradient with respect to the network output is bounded as well.

C.2 Study of $-\log(|A|)$

Here we show that the term $-\log(|A|)$ has the properties we desire.

$$\frac{\partial -\log(|A|)}{\partial A} = A^{-1T} = A^{-1} \quad (80)$$

The first step follows from Bishop Appendix C[[1](#)]. The second step comes from the fact that A is symmetric.

Bounded gradients Let D and V be the eigenvalue decomposition of B .

By using equation 38 we get

$$\left\| \frac{\partial \log(|A|)}{\partial B} \right\|_F = \|V^T (f(D)^{-1} \circ K(D, r)) V\|_F \quad (81)$$

$$= \sum_{i=0}^d 1/r(\lambda_i) \circ r'(\lambda_i) \quad (82)$$

$$\leq \sum_{i=0}^d 1/\theta \quad (83)$$

$$= \frac{d}{\theta} \quad (84)$$

Convexity We will show that the method is convex when all eigenvalues are larger than θ . i.e. when g is an identity mapping.

Since this part of the loss does not depend on v it is sufficient to prove that the loss is convex w.r.t. A .

For this part we will use a flattening function $f: \mathbb{R}^{d^2} \rightarrow \mathbb{R}^{d \times d}$ such that $f(v)_{n,m} = v_{d*(n-1)+m}$.

We will study $-\log(|f(a)|)$ and prove its convexity w.r.t. v . we will use $f(a) = A$ to simplify notation.

$$H_{d*(i-1)+j, d*(k-1)+l} = \frac{\partial^2 -\log(|f(a)|)}{\partial f(a)_{i,j} \partial f(a)_{k,l}} \quad (85)$$

$$= \frac{\partial f(a)^{-1}_{i,j}}{\partial f(a)_{k,l}} \quad (86)$$

$$= (A^{-1} \frac{\partial A}{\partial A_{k,l}} A^{-1})_{i,j} \quad (87)$$

$$= \mathbf{e}_i^T A^{-1} \mathbf{e}_k \mathbf{e}_l^T A^{-1} \mathbf{e}_j \quad (88)$$

$$= A_{i,k}^{-1} A_{l,j}^{-1} \quad (89)$$

We will show that this matrix is postive definite. Shorthand $f(x) = X$ note X needs to be positive definite.

$$\mathbf{x}^T H \mathbf{x} = \sum_{n,m=0}^d \sum_{i,j=0}^d \mathbf{x}_{d(n-1)+m} H_{d(n-1)+m,d(i-1)+j} \mathbf{x}_{d(i-1)+j} \quad (90)$$

$$= \sum_{n,m=0}^d \sum_{i,j=0}^d X_{n,m} A_{i,n}^{-1} A_{j,m}^{-1} X_{i,j} \quad (91)$$

$$= \sum_{m=0}^d X A^{-1} X A^{-1}_{m,m} \quad (92)$$

$$= \text{tr}(X A^{-1} X A^{-1}) \quad (93)$$

$$= \text{tr}(U^{-1} \hat{D} U U^{-1} \hat{D} U) \quad (94)$$

$$= \text{tr}(\hat{D} \hat{D} U^{-1} U) \quad (95)$$

$$= \sum_{n=0}^d \hat{D}_{i,i}^2 \geq 0 \quad (96)$$

We define U and \hat{D} by $U^T \hat{D} U = X A^{-1}$ such that U is ON and \hat{D} is diagonal.

The last step relies on the fact that the eigenvalues of $X A^{-1}$ are real. We will show this in the following lemma.

Lemma: Eigenvalues for multiplication of real symmetric matrices. This lemma and proof is very similar to the discussion here [14]. For two symmetric real matrices A and B where A is also positive definite then the eigenvalues of AB are real.

Proof Since A is symmetric and real there exist an eigenvalue decomposition $A = V^T D V$. Where D is diagonal, real with an inverse while V is ON. Then $AB = V^T D V B$ Then reparameterize B as $B = V^T X V$. X will still be symmetric ($X = V B V^T = V B^T V^T = V V^T X^T V V^T = X^T$) Therefore $AB = V^T D X V$, since a basis change does not change the eigenvalues AB will have the same eigenvalues as DX .

Assume d and \mathbf{v} is a pair of eigenvalues and eigenvectors of DX .

$$d \mathbf{v}^* D^{-1} \mathbf{v} = \mathbf{v}^* D^{-1} D X \mathbf{v} = \mathbf{v}^* X \mathbf{v} = \mathbf{v}^* X^* \mathbf{v} = \mathbf{v}^* X^* D^* D^{-1} \mathbf{v} = d^* \mathbf{v}^* D^{-1} \mathbf{v} \quad (97)$$

Step 1 is based on $d \mathbf{v} = D X \mathbf{v}$. Step 2 is based on $D^{-1} D = I$. Note that D^{-1} exists since A is positive definite. Step 3 is based on $X^* = X^T = X$ since X is real and symmetric. Step 4 is based on $D^{-1} D = I$ and $D^* = D$ since D is real. Step 5 is done by $\mathbf{v}^* X^* D^* = (D X \mathbf{v})^*$ since D and therefore D^{-1} is positive definite we know that $\mathbf{v}^* D^{-1} \mathbf{v} = \sum_{i=0}^d |\mathbf{v}_i|^2 / D_{i,i} > 0$. Since all $D_{i,i} > 0$. If we divide the first and last expression by this number we get $d = d^*$ and therefore d is real. This concludes the proof \square .

We use the previous lemma and conclude that our function is convex when the remapping is an identity mapping, i.e. for the set where all eigenvalues of A are larger than θ .

Bounded Hessians: If $\|\mathbf{x}\|_2 = 1$ then $\|X\|_F = 1$ and then $\|X A^{-1}\|_F \leq \|X\|_F \|A^{-1}\|_F = \|A^{-1}\|_F$

$$\text{tr}(X A^{-1} X A^{-1}) = \langle X A^{-1} \rangle^T, X A^{-1} \rangle_F \leq \|X A^{-1}\|_F^2 \leq \|A^{-1}\|_F^2 \leq \frac{d}{\theta^2} \quad (98)$$

Where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product.

We have now showed that this part of the loss has bounded gradients everywhere and that it is convex with bounded Hessians where eigenvalues are larger than θ .

C.3 Study of $h_1(\|A\mathbf{x} - \mathbf{v}\|_2)$

In this section we show that $h(\|A\mathbf{x} - \mathbf{v}\|_2)$ has the desired properties. i.e. convex respect to A and \mathbf{v} in the region where all eigenvalues of A are larger than θ , bounded Hessians for the same region and bounded gradients.

C.3.1 Properties in region $\|A\mathbf{x} - \mathbf{v}\|_2 < 1$

Here we will show that we have the desired properties in this region. If $\|A\mathbf{x} - \mathbf{v}\|_2 < 1$ then this term is

$$J = \frac{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})}{2} = \frac{\mathbf{x}^T A^T A \mathbf{x} - 2\mathbf{v}^T A \mathbf{x} + \mathbf{v}^T \mathbf{v}}{2} \quad (99)$$

We have

$$\frac{\partial J}{\partial v_i} = v_i - (\mathbf{A}\mathbf{x})_i \implies \frac{\partial J}{\partial \mathbf{v}} = (\mathbf{v} - \mathbf{A}\mathbf{x}) \quad (100)$$

$$\frac{\partial^2 J}{\partial v_i \partial v_j} = I(i = j) \quad (101)$$

$$\frac{\partial^2 J}{\partial v_i \partial A_{k,l}} = I(k = i) \mathbf{x}_l \quad (102)$$

$$\frac{\partial J}{\partial A_{i,j}} = \mathbf{x}_j (\mathbf{A}\mathbf{x} - \mathbf{v})_i \quad (103)$$

$$\frac{\partial^2 J}{\partial A_{i,j} \partial A_{k,l}} = I(k = i) \mathbf{x}_j \mathbf{x}_l \quad (104)$$

We now know the Hessian, we will use a flattening function $f: \mathbb{R}^d \times \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d+d^2}$

$$f(\mathbf{c}, B)_i = \begin{cases} \mathbf{c}_i & \text{if } i \leq d \\ B_{\lfloor i/d \rfloor, ((i \% d) + 1)} & \text{otherwise} \end{cases} \quad (105)$$

Where $\%$ indicates the remainder function.

$$f(\mathbf{c}, B)^T H f(\mathbf{c}, B) \quad (106)$$

$$= \sum_{(i,j)} I(i = j) \mathbf{c}_i^T \mathbf{c}_i + 2 \sum_{(i,k,l)} \mathbf{c}_i I(k = i) \mathbf{x}_l B_{k,l} + \sum_{i,j,k,l} B_{i,j} B_{k,l} I(k = i) \mathbf{x}_j \mathbf{x}_l \quad (107)$$

$$= \mathbf{c}^T \mathbf{c} + 2\mathbf{c}^T B \mathbf{x} + \mathbf{x}^T B^T B \mathbf{x} \quad (108)$$

$$= \|(B\mathbf{x} + \mathbf{c})\|_2^2 \geq 0 \quad (109)$$

Therefore the function is convex in this region. By maximizing B and \mathbf{c} such that $\|f(B, \mathbf{c})\|_2 = 1$ we find that the 2 norm of \mathbf{H} is $(\|\mathbf{x}\|_2^2 + 1)$. We can compute the Frobenius norm from its definition and sum and realize that $\|H\|_F = \sqrt{d * (\|\mathbf{x}\|_2^4 + \|\mathbf{x}\|_2^2 + 1)} = \sqrt{d}(\|\mathbf{x}\|_2^2 + 1)$ Where d is the dimensionality of \mathbf{x} .

In this region the gradients are bounded by

$$\sqrt{\|\mathbf{v} - \mathbf{A}\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^2 \|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2^2} < \sqrt{\|\mathbf{x}\|_2^2 + 1} \quad (110)$$

C.3.2 Properties in region $\|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2 > 1$

For this region the term turns into

$$J = \sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})} - 1/2$$

We will now compute gradients and Hessians.

$$\frac{\partial J}{\partial v_i} = \frac{(\mathbf{A}\mathbf{x} - \mathbf{v})_i}{\sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})}} \quad (111)$$

$$\frac{\partial^2 J}{\partial v_i \partial v_j} = \frac{I(i=j) \sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})} - \frac{(\mathbf{A}\mathbf{x} - \mathbf{v})_i (\mathbf{A}\mathbf{x} - \mathbf{v})_j}{\sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})}}}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^{3/2}} \quad (112)$$

$$= \frac{I(i=j) (\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}) - (\mathbf{A}\mathbf{x} - \mathbf{v})_i (\mathbf{A}\mathbf{x} - \mathbf{v})_j}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^2} \quad (113)$$

$$\frac{\partial^2 J}{\partial v_i \partial A_{k,l}} = \frac{I(k=i) \mathbf{x}_l \sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})} - \frac{(\mathbf{A}\mathbf{x} - \mathbf{v})_i \mathbf{x}_l (\mathbf{A}\mathbf{x} - \mathbf{v})_k}{\sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})}}}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^{3/2}} \quad (114)$$

$$= \mathbf{x}_l \frac{I(k=i) (\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}) - (\mathbf{A}\mathbf{x} - \mathbf{v})_i (\mathbf{A}\mathbf{x} - \mathbf{v})_k}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^2} \quad (115)$$

$$\frac{\partial J}{\partial A_{i,j}} = \frac{\mathbf{x}_j (\mathbf{A}\mathbf{x} - \mathbf{v})_i}{\sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})}} \quad (116)$$

$$\frac{\partial^2 J}{\partial A_{i,j} \partial A_{k,l}} = \mathbf{x}_j \mathbf{x}_l \frac{I(i=k) \sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})} - \frac{(\mathbf{A}\mathbf{x} - \mathbf{v})_i (\mathbf{A}\mathbf{x} - \mathbf{v})_k}{\sqrt{(\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v})}}}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^{3/2}} \quad (117)$$

$$= \mathbf{x}_j \mathbf{x}_l \frac{I(i=k) (\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}) - (\mathbf{A}\mathbf{x} - \mathbf{v})_i (\mathbf{A}\mathbf{x} - \mathbf{v})_k}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^2} \quad (118)$$

Now we see that the norm of the gradient is $\sqrt{\|\mathbf{x}\|_2^2 + 1}$

We use the flattening function again

$$f(\mathbf{c}, B)^T H f(\mathbf{c}, B) = \quad (119)$$

$$= \frac{1}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^2} \left(\sum_{i,j} \mathbf{c}_i \mathbf{c}_j (I(i=j) \|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2^2 \right. \quad (120)$$

$$- (\mathbf{A}\mathbf{x} - \mathbf{v})_i (\mathbf{A}\mathbf{x} - \mathbf{v})_j) \quad (121)$$

$$+ 2 \sum_{i,k,l} \mathbf{c}_i B_{k,l} (I(i=k) \mathbf{x}_l (\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}) - (\mathbf{A}\mathbf{x} - \mathbf{v})_i \mathbf{x}_l (\mathbf{A}\mathbf{x} - \mathbf{v})_k) \quad (122)$$

$$+ \sum_{i,j,k,l} B_{i,j} B_{k,l} \mathbf{x}_l \mathbf{x}_j (I(k=i) \|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2^2 - (\mathbf{A}\mathbf{x} - \mathbf{v})_i (\mathbf{A}\mathbf{x} - \mathbf{v})_k)) \quad (123)$$

$$= \frac{\|\mathbf{c}\|_2^2 + 2\mathbf{c}^T B\mathbf{x} + \mathbf{x}^T B^T B\mathbf{x}}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))} \quad (124)$$

$$- \frac{(\mathbf{c}^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^2 + 2\mathbf{c}^T (\mathbf{A}\mathbf{x} - \mathbf{v}) (\mathbf{A}\mathbf{x} - \mathbf{v})^T B\mathbf{x} + \mathbf{x}^T B^T (\mathbf{A}\mathbf{x} - \mathbf{v}) (\mathbf{A}\mathbf{x} - \mathbf{v}) B\mathbf{x}}{((\mathbf{A}\mathbf{x} - \mathbf{v})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^2} \quad (125)$$

$$= \frac{\|\mathbf{c} + B\mathbf{x}\|_2^2}{\|(\mathbf{A}\mathbf{x} - \mathbf{v})\|_2^2} - \frac{((\mathbf{c} + B\mathbf{x})^T (\mathbf{A}\mathbf{x} - \mathbf{v}))^2}{\|(\mathbf{A}\mathbf{x} - \mathbf{v})\|_2^4} \quad (126)$$

$$\geq \frac{\|\mathbf{c} + B\mathbf{x}\|_2^2}{\|(\mathbf{A}\mathbf{x} - \mathbf{v})\|_2^2} - \frac{\|\mathbf{c} + B\mathbf{x}\|_2^2 \|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2^2}{\|(\mathbf{A}\mathbf{x} - \mathbf{v})\|_2^4} \quad (127)$$

$$= 0 \quad (128)$$

The second to last step is from Cauchy's inequality. This concludes the proof that the function has a positive semidefinite Hessian in both regions.

We can also notice that the Hessian has eigenvalues of magnitude less than $(1 + \|\mathbf{x}\|_2^2) / \|\mathbf{A}\mathbf{x} - \mathbf{v}\|_2^2 \leq (1 + \|\mathbf{x}\|_2^2)$. The Frobenius norm of the Hessian is $\sqrt{d-1}(1 + \|\mathbf{x}\|_2^2)$.

Finally we notice that $h(\mathbf{x})$ is continuous with continuous gradients. Therefore $h(\|\mathbf{A}\mathbf{x} - \mathbf{v}\|)$ will also be continuous with continuous gradients w.r.t \mathbf{v} and \mathbf{A} .

If we consider two points (A_1, \mathbf{v}_1) and (A_2, \mathbf{v}_2) and consider the function $p(\lambda) = h(\|(\lambda A_1 + (1-\lambda)A_2)\mathbf{x} - (\lambda \mathbf{v}_1 + (1-\lambda)\mathbf{v}_2)\|_2)$. Then this function will have a positive second derivative almost everywhere and at the place where the second derivative is undefined the derivative is continuous. Therefore this function is convex. Therefore our function is convex for every line segment. Therefore the function is convex for the convex set where all eigenvalues of \mathbf{A} are larger than θ .

D Extra tables

The numbers we report for this section are based on running the same experiment 5 times with different random seeds. The number we report is the mean of these runs. The value after the \pm sign is the empirical standard deviation of these 5 runs.

Table 5: Ablation of different pretraining datasets for WFLW. The bottom two rows indicate that using pretrained imagenet weights give significant improvements over random initialization. The top two lines indicate that pretraining on a face dataset gives a small improvement in performance. All runs in this table use a resnet101 backbone with a convolution instead of average pooling at the end. The loss used is our Huber loss with v parameterization.


Pretrain dataset	epochs	NME	NLL
300W-LP[	200	4.70 ± 0.03	-344.6 ± 1.60
ImageNet	200	4.76 ± 0.06	-355.4 ± 3.30
ImageNet	50	4.91 ± 0.04	-355.8 ± 0.79
None	50	5.31 ± 0.04	-342.2 ± 1.20

Table 6: Ablation for test time augmentation (TTA) for WFLW. Using probabilistic TTA significantly improves performance compared to no TTA. All runs use resnet101 backbone with convolution instead of average pooling at the end. The methods were trained for 200 epochs. We use our loss with the v parameterization.



Pretraining dataset	TTA	NME
300W-LP[	✓	4.58 ± 0.02
300W-LP[	✗	4.70 ± 0.03
ImageNet	✓	4.62 ± 0.04
ImageNet	✗	4.76 ± 0.06

Table 7: Comparison when training for 200 epochs compared to 50 on WFLW. When training for longer the estimated position performance continues to increase for longer than the NLL. We use resnet101 backbone and our loss with v parameterization. Models are pretrained on Imagenet.

Epochs	NME	NLL
200	4.76 ± 0.06	-355.4 ± 3.30
50	4.91 ± 0.04	-355.8 ± 0.79

Table 8: Ablation of fusion type for mpii. Performance difference is small and probably not significant. Models trained for 50 epochs using resnet101 as backbone with our loss using a v parameterization.

Fusion type	PCKh@0.5
probabilistic	85.0 ± 0.1
mean	84.8 ± 0.1

Table 9: Network architecture ablation for WFLW. ResNet18 performs worse than the other two architectures. ResNet101 and ResNet50 has similar performance.

Network	NME (\downarrow)	NLL (\downarrow)
ResNet101	4.91 ± 0.04	-355.8 ± 0.79
ResNet50	4.89 ± 0.02	-357.2 ± 0.47
ResNet18	5.01 ± 0.02	-351.0 ± 0.75

Table 10: Comparison average pooling at end versus using channelwise convolutions. Experiment shows that using channelwise convolutions instead of average pooling significantly improve performance. Models use resnet101 backbone, trained for 50 epochs using our loss with v parameterization.

Average pooling at end	NME	NLL
\times	4.91 ± 0.04	-355.8 ± 0.79
\checkmark	5.25 ± 0.02	-336.5 ± 0.40

E MLE of multiple multivariate Huber distribution predictions

For many applications there will be multiple estimates of the target position. For example one could have multiple views of a person and with our approach it would be possible to generate a multivariate Huber distribution from each view, creating multiple estimates of each landmark. Unfortunately, the Huber distribution is not closed under multiplication, unlike the normal distributions. However, we have created an efficient method which is based on the majorize/minimize method for quadratic functions. For the special case $\delta = 0$ this method would turn into Weiszfeld’s algorithm [14].

We want to find the maximum likelihood point given n independent multi-variate Huber distributions. Let each independent estimate of \mathbf{y} be parameterized by (v_i, A_i) then

$$p(\mathbf{y}) \propto \prod_{i=1}^n \exp(-h_{\delta}(\|A_i \mathbf{y} - v_i\|)) \quad (129)$$

and the optimal \mathbf{y} is found from:

$$\arg \max_{\mathbf{y}} p(\mathbf{y}) = \arg \min_{\mathbf{y}} \sum_{i=1}^n h_{\delta}(\|A_i \mathbf{y} - v_i\|) = \arg \min_{\mathbf{y}} \sum_{i=1}^n g_i(\mathbf{y}) \quad (130)$$

This optimization problem can be solved with a Majorize-Minimization (MM) procedure. If

$\mathbf{y}^{(t)}$ is the current estimate for the optimal \mathbf{y} then a tight quadratic majorizer for each $g_i(\mathbf{y})$ is:

$$q_i(\mathbf{y} \mid \mathbf{y}^{(t)}) = \begin{cases} \|A_i \mathbf{y} - \mathbf{v}_i\|^2 / 2 & \text{if } \|A_i \mathbf{y}^{(t)} - \mathbf{v}_i\| < \delta \\ \frac{\delta \|A_i \mathbf{y} - \mathbf{v}_i\|^2}{2 \|A_i \mathbf{y}^{(t)} - \mathbf{v}_i\|} + \frac{\delta \|A_i \mathbf{y}^{(t)} - \mathbf{v}_i\| - \delta^2}{2} & \text{otherwise} \end{cases} \quad (131)$$

It is then simple to majorize $\sum_{i=1}^n g_i(\mathbf{y})$ with

$$q(\mathbf{y} \mid \mathbf{y}^{(t)}) = \sum_{i=1}^n q_i(\mathbf{y} \mid \mathbf{y}^{(t)}) \quad (132)$$

By iteratively solving $\mathbf{y}^{(t+1)} = \arg \min_{\mathbf{y}} q(\mathbf{y} \mid \mathbf{y}^{(t)})$, we converge to the desired maximum likelihood estimate solution. Since $q(\mathbf{y} \mid \mathbf{y}^{(t)})$ is a quadratic function with respect to \mathbf{y} finding the minima for each step is easy.

G Visualizations of L_2 multivariate Huber pdf

This section presents visualizations of the L_2 multivariate Huber distribution to aid understanding the effect of the parameters on the shape, spread and effective support of the distribution. The Λ parameter plays a similar role in the shape of the distribution as in a Gaussian distribution. The δ parameter controls the tail behaviour of the distribution and its spread given the orientation defined by Λ . Crucially, the parameters δ and Λ can be independently set to change the spread of the distribution. This means that even when δ is kept fixed one can still adapt the distribution's support via Λ to down-weight outliers in our loss. A less drastic change in Λ is needed for our Huber distribution, given a reasonable value of δ , to adapt to outliers than for a Gaussian distribution.

In the following figures it is assumed each distributions shown has zero mean vector. Each plot shows the iso-probability contours of the distribution marking the .005, .05, .2, .35, .5, .65, .8, .95 and .99 percentiles. The shading in each ring is proportional to log of the mean probability of the distribution in that region. The scaling - applied to the spatial and shading components - is constant across the plots within a figure.

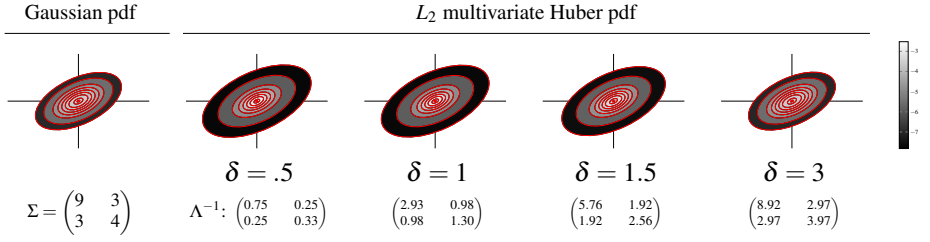


Figure 4: **Comparison of L_2 Huber distributions and a bivariate Gaussian distribution all with the same second-order moment matrix.** Leftmost plot: The bivariate normal distribution whose covariance matrix, Σ , by definition equals $E[\mathbf{X}^2]$. Other plots: Each plot shows a L_2 multivariate Huber distribution whose second order moment matrix equals that of the distribution shown in the leftmost plot. For a Huber distribution $E[\mathbf{X}^2] = \alpha(\delta)\Lambda^{-1}$. The parameters defining the shown Huber distributions are given under the plot. As δ increases: 1) The parameter matrix Λ^{-1} changes, entries increase in magnitude, to keep $E[\mathbf{X}^2]$ fixed and the spread of distribution decreases. 2) The distribution increasingly resembles a Gaussian distribution. 3) Less of the probability mass of the pdf is contained in the tails.

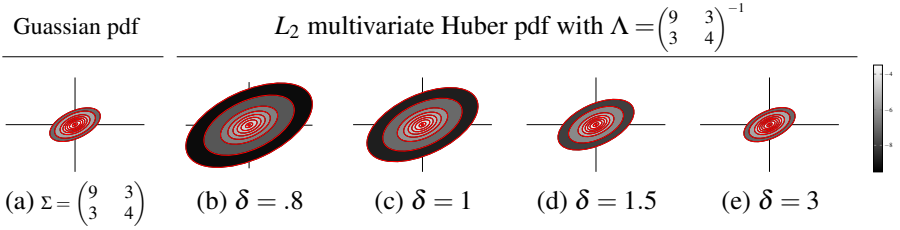


Figure 5: **Comparison of L_2 Huber distributions with the same Λ but different δ parameter.** (a) The bivariate normal distribution with covariance matrix Σ . (b-e) Each plot shows a L_2 multivariate Huber distribution with the same Λ parameter but different δ . As δ increases: 1) The spread of distribution decreases. 2) The distribution increasingly resembles a Gaussian distribution. 3) Less of the probability mass of the pdf is contained in the tails.

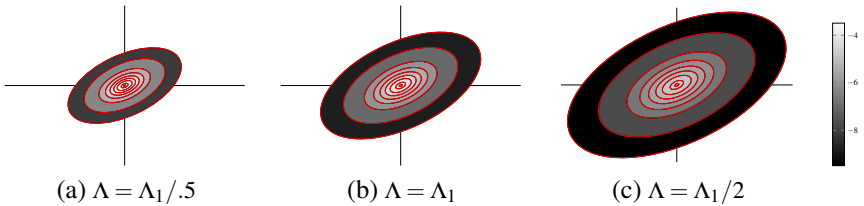


Figure 6: **Comparison of L_2 Huber distributions with the same $\delta = 1$ parameter but different Λ .** By decreasing the magnitude of values in $\Lambda_1 = \begin{pmatrix} 9 & 3 \\ 3 & 4 \end{pmatrix}^{-1}$ by the same factor one can increase the spread of the distribution.