Machine Learning in Computer Vision: Some Statistics; Convexity; Robustness

Oren Freifeld
Computer Science, Ben-Gurion University

April 25, 2018
1 Convexity

2 Linear Model, Least Squares and Weighted Least Squares

3 Outliers, “Wrong” Assumptions, and Robustness

4 Some Statistics and M-Estimators
Definition (convex function from $I \subset \mathbb{R}$ to $\mathbb{R}$)

Let $I \subset \mathbb{R}$ be an interval. $f : I \to \mathbb{R}$ is called convex if, $\forall x_1, x_2 \in I$, $\forall t \in (0, 1)$,

$$f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2)$$

and it is called strictly convex if, $\forall x_1 \neq x_2 \in I$, $\forall t \in (0, 1)$,

$$f(tx_1 + (1 - t)x_2) < tf(x_1) + (1 - t)f(x_2)$$

- convex function: a chord connecting two values of $f$ is always not below the graph of $f$.
- strictly convex function: a chord connecting two values of $f$ is always above the graph of $f$.

Example

$f(x) = |x|$ is convex. $f(x) = x^2$ is strictly convex.
Convexity

Fact
A differentiable $f : I \to \mathbb{R}$ is convex $\iff$

$$f(x) \geq f(y) + f'(y)(x - y) \quad \forall x, y \in I$$

and it is strictly convex $\iff$

$$f(x) > f(y) + f'(y)(x - y) \quad \forall x, y \in I, x \neq y$$

- a differentiable convex function $f$ is not below its tangents.
- a strictly differentiable convex function $f$ is above its tangents.

Fact
$f$ is convex and differentiable & $f'(x) = 0 \Rightarrow x$ is a global minimum of $f$. 

Fact
$f''$ exists & $f'' \geq 0$ (resp. $f'' > 0$) on $I \Rightarrow f$ is convex (strictly convex).
Convexity

Fact
A differentiable $f : I \to \mathbb{R}$ is convex $\iff$

$$f(x) \geq f(y) + f'(y)(x - y) \quad \forall x, y \in I$$

and it is strictly convex $\iff$

$$f(x) > f(y) + f'(y)(x - y) \quad \forall x, y \in I, x \neq y$$

- a differentiable convex function $f$ is not below its tangents.
- a strictly differentiable convex function $f$ is above its tangents.

Fact
$f$ is convex and differentiable & $f'(x) = 0 \Rightarrow x$ is a global minimum of $f$.

Fact
$f''$ exists & $f'' \geq 0$ (resp. $f'' > 0$) on $I \Rightarrow f$ is convex (strictly convex).
Definition (convex combination)

\[ \sum_{i=1}^{m} \alpha_i x_i, \text{ a linear combination of } \{x_i\}_{i=1}^{m} \subset \mathbb{R}^n, \text{ is called convex if all the } \alpha_i \text{'s are nonnegative and } \sum_{i=1}^{m} \alpha_i = 1. \]

Definition (convex set)

\( S \subset \mathbb{R}^n \) is called convex if it is closed under convex combinations.
Definition (convex function from a convex subset of $\mathbb{R}^n$ to $\mathbb{R}$)

Let $S \subset \mathbb{R}^n$ be convex. $f : S \to \mathbb{R}$ is called convex if,
\[
\forall x_1, x_2 \in S, \forall t \in (0, 1),
\]
\[
f(t x_1 + (1 - t) x_2) \leq t f(x_1) + (1 - t) f(x_2)
\]
and it is called strictly convex if, $\forall x_1 \neq x_2 \in I, \forall t \in (0, 1),$
\[
f(t x_1 + (1 - t) x_2) < t f(x_1) + (1 - t) f(x_2)
\]
Fact

Let $S \subset \mathbb{R}^n$ be convex. A differentiable $f : S \rightarrow \mathbb{R}$ is convex $\iff$

$$f(x) \geq f(y) + \nabla f(y)(x - y) \quad \forall x, y \in S$$

and it is strictly convex $\iff$

$$f(x) > f(y) + \nabla f(y)(x - y) \quad \forall x, y \in S, x \neq y$$

Fact

$f$ is convex and differentiable & $\nabla f(x) = 0 \Rightarrow x$ is a global minimum of $f$. 
Definition (Hessian)

the Hessian matrix of a twice-differentiable $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is

$$
\begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
\end{bmatrix}
$$

(1)
Fact

A twice-differentiable $f : S \to \mathbb{R}^n$ is convex (resp. strictly convex) $\iff$ its Hessian matrix is SPSD (SPD) on the interior of $S$. 
**Fact**

A weighted sum, with nonnegative weights, of convex functions is convex. A weighted sum, with nonnegative weights, of strictly convex functions is strictly convex.

**Example**

If \((x_i)_{i=1}^n \subset \mathbb{R}\) and \((w_i)_{i=1}^n \subset \mathbb{R}^{>0}\) (and the \(w_i\)'s do not depend on the \(x_i\)'s) then

\[
f : \mathbb{R} \rightarrow \mathbb{R} \quad f(\mu) = \sum_{i=1}^{N} w_i (\mu - x_i)^2
\]

is strictly convex while

\[
g : \mathbb{R} \rightarrow \mathbb{R} \quad g(\mu) = \sum_{i=1}^{N} w_i |\mu - x_i|
\]

is convex.
Convex Optimization

- Convex functions are “easy” to optimize.
- Sometimes they have closed-form solutions for their minimizer(s).

Exercise

\[ f : \mathbb{R} \rightarrow \mathbb{R} \quad f : x \mapsto ax^2 + bx + c \quad a \in \mathbb{R}_{>0} \quad b, c \in \mathbb{R} \]

Find \( \arg \min_x f(x) \) (hint: high school)

- But, more importantly, even when they have no closed-form minimizers it is still “easy”.
- The bible: Boyd and Vandenberghe’s Convex Optimization (available online). Also check out Boyd’s video lectures at Stanford’s YouTube channel.
The real take-home message: while you should learn the basics of convex optimization, unless you want to specialize in this field, the smart thing to do (for CV/ML researchers and practitioners) is to learn how to:

- recognize the problem is convex or
- transform a non-convex problem to a convex one
- approximate a non-convex problem using a convex one
- solve a non-convex problem by iterating between two convex subproblems.

and then be content with being a user of convex optimization.

E.g., this is how Convex Optimization is taught by Boyd at Stanford.
Linear Model

- Measurements: \( \{y_{i=1}^N\} \subset \mathbb{R}^d \)
- \( H_i \) is a known matrix, possibly dependent on \( i \).
- Residual of a linear model:

\[
\begin{align*}
\begin{bmatrix}
\mathbf{r}_i \\
\end{bmatrix}
\begin{bmatrix}
\mathbf{y}_i \\
\end{bmatrix}
\end{align*}
\]

\[
= \begin{bmatrix}
\mathbf{H}_i \\
\end{bmatrix}
\begin{bmatrix}
\mathbf{\theta} \\
\end{bmatrix}
\]

(2)

- When we discussed optical flow for gray-scale images, \( d \) was 1.
Least-Squares Estimation in a Linear Model

\[ \hat{\theta}_{\text{LS}} \triangleq \arg \min_{\theta} \sum_{i=1}^{N} \| r_i \|^2 \]

\[ H \triangleq \begin{bmatrix} H_1^T & \cdots & H_N^T \end{bmatrix}^T \in \mathbb{R}^{(Nd) \times k} \]

\[ y \triangleq \begin{bmatrix} y_1^T & \cdots & y_N^T \end{bmatrix}^T \in \mathbb{R}^{(Nd) \times 1} \]

\[ r \triangleq \begin{bmatrix} r_1^T & \cdots & r_N^T \end{bmatrix}^T = H\theta - y \in \mathbb{R}^{(Nd) \times 1} \] (3)

The cost function, from \( \mathbb{R}^k \) to \( \mathbb{R}_{>0} \), is convex; it’s strictly convex if \( \text{rank}(H) = k \).

A minimizer satisfies

\[ H^T H \hat{\theta}_{\text{LS}} = H^T y \] (4)

and is unique if \( \text{rank}(H) = k \); i.e., if \( H^T H \) is invertible.
Weighted Least-Squares

\[ \hat{\theta}_{WLS} \triangleq \arg \min_{\theta} \sum_{i=1}^{N} w_i \| r_i \|^2 \]  
\[ \left\| W^{1/2}(H\theta - y) \right\|^2 \]  
\[ W = \text{diag}(w_1, \ldots, w_1, w_2, \ldots, w_2, w_N, \ldots, w_N) \]  
\[ W^{1/2} = \text{diag}(\sqrt{w_1}, \ldots, \sqrt{w_1}, \sqrt{w_2}, \ldots, \sqrt{w_2}, \sqrt{w_N}, \ldots, \sqrt{w_N}) \]  
\[ H^T W H \hat{\theta}_{WLS} = H^T W y \]
Least-Squares Estimation is not Robust

Least Squares

Also true for weighted least squares.
Robust Least-Squares

- Measurements: \( \{ \mathbf{y}_i^{N}_{i=1} \} \subset \mathbb{R}^d \)
- Residual: \( \mathbf{r}_i = H_i \mathbf{\theta} - \mathbf{y}_i \)
- Want
  \[
  \arg \min_{\mathbf{\theta}} \sum_i \rho(\|\mathbf{r}_i\|)
  \]
  where \( \rho : \mathbb{R} \rightarrow \mathbb{R}_{>0} \) is differentiable
- \( \psi(x) = \frac{d \rho(x)}{dx} \) is called the influence function.
- If \( \rho \) is of the form \( x \mapsto x^2 \) then this is least squares.
Robust Error Function

Example (Geman-McClure robust error function)

\[ \rho(x) = \frac{x^2}{x^2 + \sigma^2} \]
Influence Function: Derivative of the Error Function

Example (Derivative of the Geman-McClure robust error function)

\[
\psi(x) = \frac{d}{dx} \rho(x) = \frac{2\sigma^2 x}{(x^2 + \sigma^2)^2}
\]
Definition (Unimodal function (AKA quasiconvext function))

Let $S \subset \mathbb{R}^n$ be a convex set. $f : S \rightarrow \mathbb{R}$ is called unimodal if all its sublevel sets

$$S_\alpha = \{x | f(x) \leq \alpha\},$$

for $\alpha \in \mathbb{R}$, are convex.

Example

Geman-McClure robust error function: $x \rightarrow x^2/(x^2 + \sigma^2)$

- A unimodal function is usually still easy to minimize. But . . .
**Fact**

A weighted sum, with nonnegative weights, of unimodal functions is usually not unimodal.

**Example**

If \((x_i)_{i=1}^n \subseteq \mathbb{R}\) and \((w_i)_{i=1}^n \subseteq \mathbb{R}_{>0}\) (and the \(w_i\)'s do not depend on the \(x_i\)'s) then

\[ f : \mathbb{R} \rightarrow \mathbb{R} \quad f(\mu) = \sum_{i=1}^{N} w_i \rho(\mu - x_i), \]

where \(\rho\) is the Geman-McClure robust error function, is usually not unimodal.
Minimizing a Nonconvex Robust Cost Function

- when $\rho(x)$ is more robust than $|x|$ then optimization is hard...
- Example for possible approaches:
  - Assuming $\rho$ is differentiable, can try gradient-based methods, possibly with graduated optimization such as
    - graduated non-convexity (minimize $(1 - t)f_{\text{convex}} + tf$); or
    - “annealing” (e.g., gradually decrease $\sigma$ in the GM robust error function)

- Iterative Reweighted Least Squares (IRLS)

Figure taken from Wikipedia
Results for a Gradient-Based Method in the Global Approach

Figure from Michael Black’s PhD, 1992

Figure 4.7: **Random Noise Example.** a) First random noise image in the sequence. b) True horizontal motion (black = −1 pixel, white = 1 pixel, gray = 0 pixels). c) True vertical motion.
Results for a Gradient-Based Method in the Global Approach

Figure from Michael Black’s PhD, 1992

Figure 4.8: Random Noise Sequence Results. a, b) Least-squares solution; horizontal and vertical components of the flow. c, d) Robust-gradient solution; horizontal and vertical components of the flow.
Results for a Gradient-Based Method in the Global Approach

*Figure from Michael Black’s PhD, 1992*

Figure 4.12: **Random Noise Sequence.** Computed flow in the case where 5 percent uniform noise is added to the second image. *a, b* Horizontal and vertical least-squares flow. *c, d* Horizontal and vertical robust flow.
Results for a Gradient-Based Method in the Global Approach

Figure from Michael Black’s PhD, 1992

Figure 4.14: Effect of robust data term, (10% uniform noise). a,b) Least-squares (quadratic) solution. c,d) Quadratic data term and robust smoothness term. e,f) Fully robust formulation.
Iterative Reweighted Least-Squares

- Idea: use $w(x) = \psi(x)/x$
- Start with an LS solution, then alternate between computing weights based on the residual errors, and a WLS solution using fixed weights from the previous iteration.
  - Initialization: solve
    $$H^T H \hat{\theta}_{\text{IRLS}}^{[0]} = H^T y$$  \hspace{1cm} (8)
  - Alternate
    1. Set $w_i^{[k]} = \frac{\psi(r_i^{[k]})}{\|r_i^{[k]}\|}$ where $r_i^{[k]} = H_i \hat{\theta}_{\text{IRLS}}^{[k]} - y_i$
    2. Solve
       $$H^T W^{[k]} H \hat{\theta}_{\text{IRLS}}^{[k+1]} = H^T W^{[k]} y$$ \hspace{1cm} (9)
Iterative Reweighted Least-Squares

\[ E_{RLS} = \sum_i \rho(\|r_i\|) \]

\[ \nabla \theta E_{RLS} = \sum_i \psi(\|r_i\|) \nabla \theta \|r_i\| \overset{\text{want}}{=} 0 \] \hspace{1cm} (10)

\[ E_{IRLS} = \sum_i w(r_i) \|r_i\|^2 \] \hspace{1cm} (11)

\[ E_{IRLS}^{[k+1]} = \sum_i w(r_i^{[k]}) \|r_i\|^2 = \sum_i \frac{\psi(\|r_i^{[k]}\|)}{\|r_i^{[k]}\|} \|r_i\|^2 \approx \sum_i \psi(\|r_i^{[k]}\|) \|r_i\| \]

\[ 0 = \nabla \theta E_{IRLS}^{[k+1]} \approx \sum_i \psi(\|r_i^{[k]}\|) \nabla \theta \|r_i\| \]
Iterative Reweighted Least-Squares

\[ E_{\text{RLS}} = \sum_i \rho(\|r_i\|) \]

\[ \nabla_\theta E_{\text{RLS}} = \sum_i \psi(\|r_i\|) \nabla_\theta \|r_i\| \xrightarrow{\text{want}} 0 \]  \hspace{1cm} (10)

\[ E_{\text{IRLS}} = \sum_i w(r_i) \|r_i\|^2 \]  \hspace{1cm} (11)

\[ E_{\text{IRLS}}^{[k+1]} = \sum_i w(r_i^{[k]}) \|r_i\|^2 = \sum_i \frac{\psi \left( \|r_i^{[k]}\| \right)}{\|r_i^{[k]}\|} \|r_i\|^2 \approx \sum_i \psi \left( \|r_i^{[k]}\| \right) \|r_i\| \]

\[ 0 = \nabla_\theta E_{\text{IRLS}}^{[k+1]} \approx \sum_i \psi \left( \|r_i^{[k]}\| \right) \nabla_\theta \|r_i\| \]
Least Squares

- True
- LS
Iterative Rewighted Least Squares, iter 1
Iterative Rewighted Least Squares, iter 2

- True
- IRLS

Apr 25, 2018 29 / 78
Iterative Rewighted Least Squares, iter 9

- True path
- IRLS path
Iterative Rewighted Least Squares, iter 12

true

IRLS
Iterative Rewighted Least Squares, iter 13
Iterative Rewighted Least Squares, iter 15

- True
- IRLS
Recall Lucas-Kanade

\[
\sum_{i=1}^{N} \left( \sum_{i=1}^{N} \left( \begin{array}{c}
\sqrt{g(x, x_i)} \nabla_x I(x_i, t) \\
H_i : 1 \times 2
\end{array} \right) \begin{bmatrix}
u(x) \\
u(x)
\end{bmatrix} + \sqrt{g(x, x_i)} I_t(x_i, t) \right) \right)^2 = \|W^{1/2} \varepsilon\|^2 = \varepsilon^T W \varepsilon =
\]

\[
W = \text{diag} \left( \begin{bmatrix} g(x, x_1) & \ldots & g(x, x_N) \end{bmatrix} \right) \quad \text{(not the same } W \text{ from IRLS)}
\]

\[
W^{1/2} = \text{diag} \left( \begin{bmatrix} \sqrt{g(x, x_1)} & \ldots & \sqrt{g(x, x_N)} \end{bmatrix} \right)
\]

\[
\begin{bmatrix} \hat{u}(x) \\
\hat{v}(x) \end{bmatrix}_{\text{WLS}} \triangleq \text{arg min}_{u(x), v(x)} \varepsilon^T W \varepsilon
\]
Robust Lucas-Kanade

\[ E_{RLK} = \sum_{i=1}^{N} \rho \left( \sqrt{g(x, x_i)} \nabla_x I(x_i, t) \left[ \begin{array}{c} u(x) \\ v(x) \end{array} \right] + \sqrt{g(x, x_i)} I_t(x_i, t) \right) \]

\[ H_i: 1 \times 2 \]

\[ \theta: 2 \times 1 \]

\[ r_i \in \mathbb{R} \]

\[ y_i \in \mathbb{R} \]

(15)
Robust Affine Lucas-Kanade

\[
\begin{align*}
\begin{bmatrix}
  u(x_i) \\
v(x_i)
\end{bmatrix} &=
\begin{bmatrix}
x_i - x & y_i - y & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & x_i - x & y_i - y & 1
\end{bmatrix}
\begin{bmatrix}
\theta_1 \\
\theta_2 \\
\theta_3 \\
\theta_4 \\
\theta_5 \\
\theta_6
\end{bmatrix} \\
&= A(x, x_i) \Delta \\
\Rightarrow \\
E_{RALK} &= \sum_{i=1}^{N} \rho 
\begin{pmatrix}
\sqrt{g(x, x_i)} \nabla_x I(x_i, t) & A(x, x_i) & \theta + \sqrt{g(x, x_i)} I_t(x_i, t)
\end{pmatrix}
\begin{bmatrix}
H_i; 1\times6 \\
r_i \in \mathbb{R}
\end{pmatrix}
\begin{bmatrix}
-y_i \in \mathbb{R}
\end{pmatrix}
\end{align*}
\]
Random Sample Consensus (RANSAC)

- A non-deterministic algorithm
- For generality, write the model as $f(x_i, \theta) = y_i$
- Stick to the least-squares formulation
- Algorithm: given data $\{x_i, y_i\}_{i=1}^N$, alternate between
  - Pick a very small random subset of the data. Its cardinality should suffice for estimating the parameters (e.g., you can’t estimate a line from a single point). Fit a least-squares model to it.
  - For each data point in the original set, compute $\|f(x_i, \hat{\theta}) - y_i\|$. Using a threshold, reject outliers.

Once a certain iteration achieves enough inliers (points not rejected), the algorithm stops.
RANSAC and Optical Flow

- For HS-type models, RANSAC is inapplicable.
- For LK-type models, it fits only if the neighborhood is sufficiently large. For example, if we try to find a single affine flow for the entire image.
Definition (sample mean for $\mathbb{R}$-valued data)

The sample mean of $\{x_i\}_{i=1}^N \subset \mathbb{R}$ is

$$\bar{x} \triangleq \frac{1}{N} \sum_{i=1}^{N} x_i \quad (18)$$

Definition (sample mean for $\mathbb{R}^n$-valued data)

The sample mean of $\{x_i\}_{i=1}^N \subset \mathbb{R}^n$ is

$$\bar{x} \triangleq \frac{1}{N} \sum_{i=1}^{N} x_i \quad (19)$$
Definition (sample mean for $\mathbb{R}$-valued data)

The sample mean of $\{x_i\}_{i=1}^{N} \subset \mathbb{R}$ is

$$
\bar{x} \triangleq \frac{1}{N} \sum_{i=1}^{N} x_i
$$

(18)

Definition (sample mean for $\mathbb{R}^n$-valued data)

The sample mean of $\{x_i\}_{i=1}^{N} \subset \mathbb{R}^n$ is

$$
\bar{x} \triangleq \frac{1}{N} \sum_{i=1}^{N} x_i
$$

(19)
Fact

The sample mean minimizes the sum of squared Euclidean distances:

\[ \bar{x} = \arg \min_{\mu \in \mathbb{R}} \sum_{i=1}^{N} (x_i - \mu)^2. \]
Example

Let $x_1, x_2 \in \mathbb{R}$. Then

$$\bar{x} = \frac{x_1 + x_2}{2} = \arg\min_{\mu \in \mathbb{R}} (x_1 - \mu)^2 + (x_2 - \mu)^2.$$  \hfill (21)
The Sample Mean as a Minimizer

Fact

The sample mean minimizes the sum of squared scaled Euclidean distances:

$$\bar{x} = \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \quad \sigma > 0.$$  \hspace{1cm} (22)

Proof.

$$\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\arg\min_{\mu \in \mathbb{R}} \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 = \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{N} (x_i - \mu)^2 = \bar{x}$$
The Sample Mean as a Minimizer

Fact

The sample mean minimizes the sum of squared scaled Euclidean distances:

$$\bar{x} = \arg \min_{\mu \in \mathbb{R}} \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \quad \sigma > 0.$$  (22)

Proof.

$$\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\arg \min_{\mu \in \mathbb{R}} \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 = \arg \min_{\mu \in \mathbb{R}} \sum_{i=1}^{N} (x_i - \mu)^2 = \bar{x}$$
Example

Let $x_1, x_2 \in \mathbb{R}$. Then

$$\bar{x} = \frac{x_1 + x_2}{2} = \arg \min_{\mu \in \mathbb{R}} \frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_2 - \mu)^2}{\sigma^2}.$$  \hspace{1cm} (23)
\[
\sigma = 2:
\]

\[
\frac{(\mu - x_1)^2}{\sigma^2}
\]

\[
\frac{(\mu - x_2)^2}{\sigma^2}
\]

both funcs

\[
\sum_i \frac{(x_i - \mu)^2}{\sigma^2}
\]

sum both funcs

sample mean
\[ \sigma = 4: \]

\[ \frac{(\mu - x_1)^2}{\sigma^2} \]

\[ \frac{(\mu - x_2)^2}{\sigma^2} \]

both funcs

\[ \sum_{i} \frac{(x_i - \mu)^2}{\sigma^2} \]

sample mean
$\sigma = 4, \ N = 3$: 

\[ \Sigma_i (x_i - \mu)^2 / \sigma^2 \]
$\sigma = 4, \ N = 4$: 

![Graph showing all 4 functions and the sum of all functions with sample mean marked.]
σ = 4, \( N = 4 \):
\( \sigma = 4, \ N = 7 \) (not robust to outliers):
The Sample Mean as a Minimizer

Fact

\[ \bar{x} = \arg\min_{\mu \in \mathbb{R}^n} \sum_{i=1}^{N} \left\| \frac{1}{\sigma} (x_i - \mu) \right\|_{\ell_2}^2. \]  

(24)

Outline of the proof:

(i) Show that, \( \nabla_\mu E(\mu) \), the gradient of \( E(\mu) = \sum_{i=1}^{N} \left\| \frac{1}{\sigma} (x_i - \mu) \right\|_{\ell_2}^2 \) w.r.t. \( \mu = [\mu_1 \ldots \mu_n]^T \), is proportional to

\[
\begin{bmatrix}
\sum_{i=1}^{N} (\mu_1 - x_{i,1}) \\
\vdots \\
\sum_{i=1}^{N} (\mu_n - x_{i,n})
\end{bmatrix}
\]

(25)

where \( x_{i,j} \) is the \( j \)-th entry of \( x_i \).

(ii) Solve \( \nabla_\mu E(\mu) = 0 \)
The Sample Mean as a Minimizer

Fact

$$\bar{x} = \arg \min_{\mu \in \mathbb{R}^n} \sum_{i=1}^{N} \left\| \frac{1}{\sigma} (x_i - \mu) \right\|_{l_2}^2.$$  \hfill (24)

Outline of the proof:

(i) Show that, $\nabla_{\mu} E(\mu)$, the gradient of $E(\mu) = \sum_{i=1}^{N} \left\| \frac{1}{\sigma} (x_i - \mu) \right\|_{l_2}^2$ w.r.t. $\mu = [\mu_1 \ldots \mu_n]^T$, is proportional to

$$\left[ \sum_{i=1}^{N}(\mu_1 - x_{i,1}) \ldots \sum_{i=1}^{N}(\mu_n - x_{i,n}) \right] \hfill (25)$$

where $x_{i,j}$ is the $j$-th entry of $x_i$.

(ii) Solve $\nabla_{\mu} E(\mu) = 0$
More generally, let $Q$ be an SPD $n$-by-$n$ matrix.

**Fact**

$$\bar{x} = \arg\min_{\mu \in \mathbb{R}^n} \sum_{i=1}^{N} \|x_i - \mu\|_Q^2$$

(26)

where

$$\|x_i - \mu\|_Q^2 = (x_i - \mu)^T Q (x_i - \mu)$$

If $Q = \sigma^{-2} I_{n \times n} \propto I_{n \times n}$ this reduces to the previous problem:

$$\arg\min_{\mu \in \mathbb{R}^n} \sum_{i=1}^{N} \left\| \frac{1}{\sigma} (x_i - \mu) \right\|_{\ell_2}^2$$
Informal Definition (Statistic)

A statistic is a function that depends only on the data.

Trivially, a function of a statistic is also a statistic.

Example

\[ S_1(x_1, \ldots, x_N; N) = \sum_{i=1}^{N} x_i \in \mathbb{R}^n \] is a statistic of \((x_1, \ldots, x_N) \subset \mathbb{R}^n\). The same goes for \(\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} S_1(x_1, \ldots, x_N; N)\).

Example

\[ S_2(x_1, \ldots, x_N; N) = \sum_{i=1}^{N} x_i x_i^T \in \mathbb{R}^{n \times n} \] is a statistic of \((x_1, \ldots, x_N) \subset \mathbb{R}^n\). The same goes for \(\frac{1}{N} S_2(x_1, \ldots, x_N)\) and \(\left(\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T\right) - \bar{x} \bar{x}^T\).
Informal Definition (Statistic)

A statistic is a function that depends only on the data.

Trivially, a function of a statistic is also a statistic.

Example

\[ S_1(x_1, \ldots, x_N; N) = \sum_{i=1}^{N} x_i \in \mathbb{R}^n \text{ is a statistic of } (x_1, \ldots, x_N) \subset \mathbb{R}^n. \]

The same goes for \( \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} S_1(x_1, \ldots, x_N; N). \)

Example

\[ S_2(x_1, \ldots, x_N; N) = \sum_{i=1}^{N} x_i x_i^T \in \mathbb{R}^{n \times n} \text{ is a statistic of } (x_1, \ldots, x_N) \subset \mathbb{R}^n. \text{ The same goes for } \frac{1}{N} S_2(x_1, \ldots, x_N) \text{ and } \left( \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \right) - \bar{x} \bar{x}^T \]
Informal Definition (Statistic)

A statistic is a function that depends only on the data.

Trivially, a function of a statistic is also a statistic.

Example

\[ S_1(x_1, \ldots, x_N; N) = \sum_{i=1}^{N} x_i \in \mathbb{R}^n \text{ is a statistic of } (x_1, \ldots, x_N) \subset \mathbb{R}^n. \]

The same goes for \( \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} S_1(x_1, \ldots, x_N; N). \)

Example

\[ S_2(x_1, \ldots, x_N; N) = \sum_{i=1}^{N} x_i x_i^T \in \mathbb{R}^{n \times n} \text{ is a statistic of } (x_1, \ldots, x_N) \subset \mathbb{R}^n. \text{ The same goes for } \frac{1}{N} S_2(x_1, \ldots, x_N) \text{ and } \left( \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \right) - \bar{x} \bar{x}^T. \]
Definition (k-th order statistic)

The k-th order statistic of \( \{x_i\}_{i=1}^N \subset \mathbb{R} \) is the k-th-smallest value among \( \{x_i\}_{i=1}^N \). It is denoted by \( x(k) \). Thus, \( x(1) \leq x(2) \leq \ldots \leq x(N) \).

Example

\[ x(1) = \min \{ x_1, \ldots, x_N \} \quad \text{and} \quad x(N) = \max \{ x_1, \ldots, x_N \}. \]

Definition (order statistics)

The ordered \( N \)-tuple of the sorted values,

\[ (x(1), x(2), \ldots, x(N)) , \]

is called the order statistics of \( \{x_i\}_{i=1}^N \).
Definition ($k$-th order statistic)

The $k$-th order statistic of $\{x_i\}_{i=1}^N \subset \mathbb{R}$ is the $k$-th-smallest value among $\{x_i\}_{i=1}^N$. It is denoted by $x_{(k)}$. Thus, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}$.

Example

$x_{(1)} = \min\{x_1, \ldots, x_N\}$ and $x_{(N)} = \max\{x_1, \ldots, x_N\}$.

Definition (order statistics)

The ordered $N$-tuple of the sorted values,

$$(x_{(1)}, x_{(2)}, \ldots, x_{(N)}),$$

is called the order statistics of $\{x_i\}_{i=1}^N$. 

(27)
Some Statistics and M-Estimators

Definition \((k\text{-th order statistic})\)

The \(k\)-th order statistic of \(\{x_i\}_{i=1}^{N} \subset \mathbb{R}\) is the \(k\)-th-smallest value among \(\{x_i\}_{i=1}^{N}\). It is denoted by \(x_{(k)}\). Thus, \(x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}\).

Example

\[
x_{(1)} = \min \{x_1, \ldots, x_N\} \quad \text{and} \quad x_{(N)} = \max \{x_1, \ldots, x_N\}.
\]

Definition (order statistics)

The ordered \(N\)-tuple of the sorted values,

\[
(x_{(1)}, x_{(2)}, \ldots, x_{(N)}),
\]

is called the order statistics of \(\{x_i\}_{i=1}^{N}\).
For simplicity, let us assume $N$ is odd.

**Definition (Sample median)**

If $N$ is odd, then the sample median of $\{x_i\}_{i=1}^{N} \subset \mathbb{R}$, is

$$x\left(\frac{N+1}{2}\right)$$

(28)

If $N$ is even, there are several different ways to define the sample median; when $N$ is large then they usually become similar.
For simplicity, let us assume $N$ is odd.

**Definition (Sample median)**

If $N$ is odd, then the sample median of $\{x_i\}_{i=1}^N \subset \mathbb{R}$, is

$$x\left(\frac{N+1}{2}\right)$$

(28)

If $N$ is even, there are several different ways to define the sample median; when $N$ is large then they usually become similar.
The Sample Median as a Minimizer

Fact

The sample median minimizes the sum of scaled $\ell_1$ distances:

$$x\left(\frac{N+1}{2}\right) = \arg \min_{m \in \mathbb{R}} \sum_{i=1}^{N} \frac{|x_i - m|}{\sigma} \quad \sigma > 0.$$  \hspace{1cm} (29)

We may take this as the definition of the sample median (in which case, we don’t need to worry if $N$ is even or odd).
$\sigma = 2$:

\[ |\mu - x_1|/\sigma \]
\[ |\mu - x_2|/\sigma \]
\[ \text{both funcs} \]
\[ \sum_{i} |x_i - \mu|/\sigma \]

Sample median
$\sigma = 4$:

- $|\mu - x_1|/\sigma$
- $|\mu - x_2|/\sigma$
- both funcs
- sum both funcs

The graphs show the function $|\mu - x_i|/\sigma$ for $\mu$ values ranging from $-30$ to $30$. The sample median is indicated by a red dot for each function.
σ = 4, N = 3:
\[ \sigma = 4, \ N = 4: \]

![Graphs showing all 4 functions and sum all functions with indicated sample median.](image)
\( \sigma = 4, \ N = 4: \)

![Graph showing relationships between \( \mu \) and \( \sigma \)]
\( \sigma = 4, \quad N = 7 \) (more robust to outliers):
Since $\ell_1$ is more robust than $\ell_2$, this interpretation in terms of optimization problems explains why the sample median is more robust than the sample mean.
Median Filtering

Replace pixel \((i, j)\) with the median of the value in, say, a 5 by 5, neighborhood around it. This is a nonlinear operation.

Images taken from Wikipedia
Trimmed Average

The median is also an extreme case of the truncated (or trimmed) average:

\[
\frac{1}{N - 2N_0} \sum_{i=N_0}^{N-N_0} x(i)
\]

(30)

We will return to the trimmed average later when we discuss Robust PCA (PCA is a dimensionality-reduction technique which we will discuss as well).
M-Estimators

More generally, estimators that are defined as minimizers of sums functions of the data are called M-estimators. These include, among other things,

- \( \sum_{i=1}^{N} \rho(r_i) \) we saw previously.

Maximum-Likelihood Estimators (MLE):

\[
\arg \max_{\theta} \prod_{i=1}^{N} p(x_i; \theta) = \arg \min_{\theta} \sum_{i=1}^{N} - \log p(x_i; \theta)
\]

MLE enjoy many desired properties and satisfy various asymptotic optimality criteria – but may suffer from outliers. Some robust estimators achieve near-optimality when there are no outliers, and suffer little in their presence.

- \( M \)-estimators can be defined even when the space is nonlinear — we will see some examples.
M-Estimators

More generally, estimators that are defined as minimizers of sums functions of the data are called M-estimators. These include, among other things,

- $\sum_{i=1}^{N} \rho(r_i)$ we saw previously.
- Maximum-Likelihood Estimators (MLE):

$$\arg \max_{\theta} \prod_{i=1}^{N} p(x_i; \theta) = \arg \min_{\theta} \sum_{i=1}^{N} - \log p(x_i; \theta)$$

MLE enjoy many desired properties and satisfy various asymptotic optimality criteria — but may suffer from outliers. Some robust estimators achieve near-optimality when there are no outliers, and suffer little in their presence.

- $M$-estimators can be defined even when the space is nonlinear — we will see some examples.
More generally, estimators that are defined as minimizers of sums of functions of the data are called M-estimators. These include, among other things,

- \( \sum_{i=1}^{N} \rho(r_i) \) we saw previously.

- Maximum-Likelihood Estimators (MLE):

\[
\arg \max_{\theta} \prod_{i=1}^{N} p(x_i; \theta) = \arg \min_{\theta} \sum_{i=1}^{N} - \log p(x_i; \theta)
\]

MLE enjoy many desired properties and satisfy various asymptotic optimality criteria — but may suffer from outliers. Some robust estimators achieve near-optimality when there are no outliers, and suffer little in their presence.

- \( M \)-estimators can be defined even when the space is nonlinear — we will see some examples.
M-Estimators

More generally, estimators that are defined as minimizers of sums functions of the data are called M-estimators. These include, among other things,

- $\sum_{i=1}^{N} \rho(r_i)$ we saw previously.
- Maximum-Likelihood Estimators (MLE):

$$\arg \max_{\theta} \prod_{i=1}^{N} p(x_i; \theta) = \arg \min_{\theta} \sum_{i=1}^{N} -\log p(x_i; \theta)$$

MLE enjoy many desired properties and satisfy various asymptotic optimality criteria – but may suffer from outliers. Some robust estimators achieve near-optimality when there are no outliers, and suffer little in their presence.

$M$-estimators can be defined even when the space is nonlinear — we will see some examples.
The Sample Mean as an MLE

Fact

The sample mean maximizes the likelihood under a Gaussian likelihood model:

$$\bar{x} = \arg \max_{\mu \in \mathbb{R}} \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2).$$  \hspace{1cm} (31)

Remark

$$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2) \, dx_1 \cdots dx_N = 1$$

$$\int_{\mathbb{R}} \mathcal{N}(x_i; \mu, \sigma^2) \, d\mu = 1 \ \forall i \text{ but } \int_{\mathbb{R}} \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2) \, d\mu \neq 1 \ \forall N > 1$$
Example

Let \((x_1, x_2) \sim \mathcal{N}(x_1; \mu, \sigma^2)\mathcal{N}(x_2; \mu, \sigma^2)\).

\[
\bar{x} = \frac{x_1 + x_2}{2} = \arg \max_{\mu} \mathcal{N}(x_1; \mu, \sigma^2)\mathcal{N}(x_2; \mu, \sigma^2)
\] (32)
\( \sigma = 2: \)

\[
(2\sigma^2\pi)^{-1/2}\exp\left(-\frac{(\mu - x_1)^2}{\sigma^2}\right)
\]

\[
(2\sigma^2\pi)^{-1/2}\exp\left(-\frac{(\mu - x_2)^2}{\sigma^2}\right)
\]

both funcs

\[
(2\pi\sigma^2)^{-n/2}\prod_i \exp\left(-0.5\frac{(x_i - \mu)^2}{\sigma^2}\right)
\]

sample mean
$$\sigma = 4$$:

$$\frac{(2\sigma^2 \pi)^{-1/2}}{\exp\left(-\frac{(\mu - x_1)^2}{2\sigma^2}\right)}$$

$$\frac{(2\sigma^2 \pi)^{-1/2}}{\exp\left(-\frac{(\mu - x_2)^2}{2\sigma^2}\right)}$$

both funcs

product of both funcs

sample mean

$$\frac{(2\pi \sigma^2)^{-n/2}}{\prod_i \exp\left(-0.5\left(x_i - \mu\right)^2 / \sigma^2\right)}$$
$\sigma = 4, \ N = 3$: 

\[
\begin{align*}
\text{all 3 funcs} & \quad \text{product all funcs} \\
\mu & \quad 1e-7 \\
\end{align*}
\]

\[
\frac{(2\pi\sigma^2)^{-n/2}}{\prod_i \exp(-0.5(x_i - \mu)^2/\sigma^2)}
\]

- sample mean
\[ \sigma = 4, \; N = 4: \]

\[ (2\pi\sigma^2)^{-n/2} \prod_i \exp\left( -0.5\left( x_i - \mu \right)^2 / \sigma^2 \right) \]

- sample mean

\[ \mu \]

\[ \mu \]
$\sigma = 4, \ N = 4$:
Some Statistics and M-Estimators

\[\sigma = 4, \ N = 7 \text{ (not robust to outliers)}:\]

\[
(2\pi\sigma^2)^{-n/2} \prod \exp \left( -0.5 \frac{(x_i - \mu)^2}{\sigma^2} \right)
\]

- sample mean
More generally, let $\Sigma$ be an SPD $n$-by-$n$ matrix.

**Fact**

The inverse of an SPD matrix is also SPD.

**Fact**

$$\bar{x} = \arg \max_{\mu \in \mathbb{R}^n} \prod_{i=1}^{N} \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp \left( -\frac{1}{2} \left\| x_i - \mu \right\|_{\Sigma^{-1}}^2 \right)$$

(33)

where

$$\left\| x_i - \mu \right\|_{\Sigma^{-1}}^2 = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu),$$

is called the Mahalanobis distance.

$|\Sigma|$ is the determinant of $\Sigma$
The Sample Median as an MLE

Fact

The sample median maximizes the likelihood under a Laplace-distribution likelihood model

\[ \bar{x} = \arg \max_{\mu \in \mathbb{R}} \prod_{i=1}^{N} f(x_i; \mu, \sigma). \quad (34) \]

where

\[ f(x; \mu, \sigma) = \frac{1}{2\sigma} \exp \left( -\frac{|x - \mu|}{\sigma} \right) \]
Some Statistics and M-Estimators

\[
(2\sigma)^{-1}\exp\left(-\frac{|\mu - x_1|}{\sigma}\right) \quad \text{(2)}
\]

\[
(2\sigma)^{-1}\exp\left(-\frac{|\mu - x_2|}{\sigma}\right) \quad \text{both funcs}
\]

\[
(2\sigma)^{-1/n}\prod_i \exp\left(-\frac{|x_i - \mu|}{\sigma}\right) \quad \text{product of both funcs}
\]

---

\[
\text{sample median}
\]

---

**C**
$\sigma = 4, N = 3$: 

\[
(2\sigma)^{-1/n} \prod_i \exp\left(-\frac{|x_i - \mu|}{\sigma}\right)
\] 

sample median
σ = 4, N = 4:

![Graph showing the relationship between μ and the probability density function for different values of σ, with the sample median highlighted. The graph is divided into two sections: one showing all 4 functions and the other showing the product of all functions. The equation for the product of all functions is also provided: $(2σ)^{-1/n} \prod_{i} \exp(-|x_i - \mu|/σ)$. Sample median is marked on the graph.](image-url)
$\sigma = 4, \ N = 4$: 

![Graph showing distribution of data with $\sigma = 4$ and $N = 4$.]
$\sigma = 4$, $N = 7$ (more robust to outliers than the Gaussian):
Definition (sample correlation matrix)
\[
\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \in \mathbb{R}^{n \times n}
\] is called the sample correlation matrix of \(\{x_i\}_{i=1}^{N} \subset \mathbb{R}^n\).

Definition (sample covariance matrix)
\[
\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T
\] is called the sample covariance matrix of \(\{x_i\}_{i=1}^{N} \subset \mathbb{R}^n\).

Fact
\[
\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T = \left( \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \right) - \bar{x}\bar{x}^T
\]
\[
\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})x_i^T = \frac{1}{N} \sum_{i=1}^{N} x_i(x_i - \bar{x})^T
\]
Definition (sample correlation matrix)

\[ \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \in \mathbb{R}^{n \times n} \text{ is called the sample correlation matrix of} \]
\[ \{x_i\}_{i=1}^{N} \subset \mathbb{R}^n \]

Definition (sample covariance matrix)

\[ \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T \text{ is called the sample covariance matrix of} \]
\[ \{x_i\}_{i=1}^{N} \subset \mathbb{R}^n \]

Fact

\[ \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T = \left( \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \right) - \bar{x} \bar{x}^T \]
\[ \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x}) x_i^T = \frac{1}{N} \sum_{i=1}^{N} x_i (x_i - \bar{x})^T \]
Fact

Both $\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T$ and $\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$ are symmetric $n \times n$ matrices and their eigenvalues are always nonnegative. $
Rightarrow$ they are SPSD. If their eigenvalues are positive, then they are also SPD.