Peeter Joot
peeterjoot@protonmail.com

UofT ece1505 convex optimization: Introduction (taught by Prof. Stark Draper)

1.1 What's this course about?

- Science of optimization.

- problem formulation, design, analysis of engineering systems.

1.2 Basic concepts

- Basic concepts. convex sets, functions, problems.

- Theory (about 40 % of the material). Specifically Lagrangian duality.

- Algorithms: gradient descent, Newton's, interior point, ...

Homework will involve computational work (solving problems, ...)

1.3 Goals

- Recognize and formulate engineering problems as convex optimization problems.

- To develop (Matlab) code to solve problems numerically.

- To characterize the solutions via duality theory

- NOT a math course, but lots of proofs.

- NOT a communications course, but lots of ... (?)

- NOT a CS course, but lots of useful algorithms.

Definition 1.1: Mathematical program

$$\min_{\mathbf{x}} F_0(\mathbf{x}) \tag{1.1}$$

---

where $\mathbf{x} = (x_1, x_2, \cdots, x_m) \in \mathbb{R}^m$ is subject to constraints $F_i : \mathbb{R}^m \to \mathbb{R}^1$

$$F_i(\mathbf{x}) \leq 0, \qquad i = 1, \cdots, m \tag{1.2}$$

The function $F_0 : \mathbb{R}^m \to \mathbb{R}^1$ is called the "objective function".

Solving a problem produces:

An optimal $\mathbf{x}^*$ is a value $\mathbf{x}$ that gives the smallest value among all the feasible $\mathbf{x}$ for the objective function F_0. Such a function is sketched in fig. 1.1.
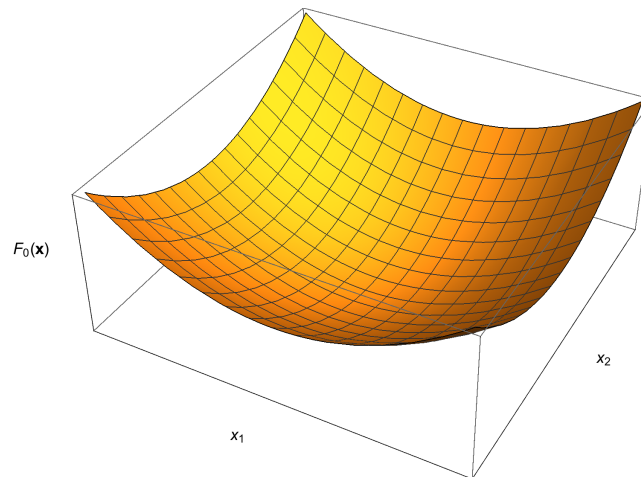

Figure 1.1: Convex objective function.

- A convex objective looks like a bowl, "holds water".

- If connect two feasible points line segment in the ? above bottom of the bowl.

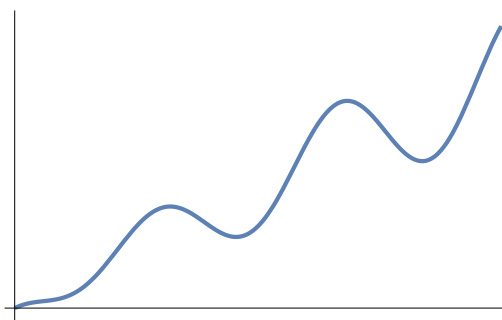A non-convex function is illustrated in fig. 1.2, which has a number of local minimums.

Figure 1.2: Non-convex (wavy) figure with a number of local minimums.

1.4 Example: Line fitting.

A linear fit of some points distributed around a line $y = ax + b$ is plotted in fig. 1.3. Here a, b are the optimization variables $\mathbf{x} = (a, b)$.
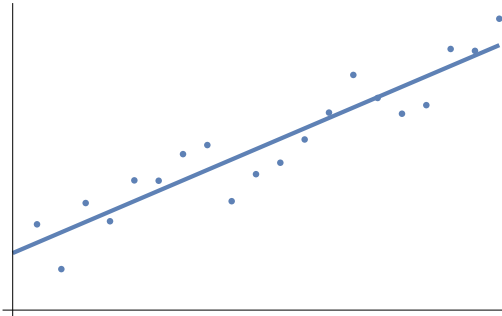

Figure 1.3: Linear fit of points around a line.

How is the solution for such a best fit line obtained?

Approach 1: Calculus minimization of a multivariable error function. Describe an error function, describing how far from the line a given point is.

$$y_i - (ax_i + b), \tag{1.3}$$

Because this can be positive or negative, we can define a squared variant of this, and then sum over all data points.

$$F_0 = \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2. \tag{1.4}$$

One way to solve (for a, b): Take the derivatives

$$\frac{\partial F_0}{\partial a} = \sum_{i=1}^{n} 2(y_i - (ax_i + b))(-x_i) = 0$$

$$\frac{\partial F_0}{\partial b} = \sum_{i=1}^{n} 2(y_i - (ax_i + b))(-1) = 0. \tag{1.5}$$

This yields

$$\sum_{i=1}^{n} y_i = \left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} 1\right) b$$

$$\sum_{i=1}^{n} x_i y_i = \left(\sum_{i=1}^{n} x_i^2\right) a + \left(\sum_{i=1}^{n} x_i\right) b. \tag{1.6}$$

In matrix form, this is

$$\begin{bmatrix} \sum x_i y_i \\ \sum y_i \end{bmatrix} = \begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}. \tag{1.7}$$

If invertible, have an analytic solution for (a^*, b^*). This is a convex optimization problem because $F(x) = x^2$ is a convex "quadratic program". In general a quadratic program has the structure

$$F(a, b) = (\cdots)a^2 + (\cdots)ab + (\cdots)b^2. \tag{1.8}$$

Approach 2: Linear algebraic formulation.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}, \tag{1.9}$$

or

$$\mathbf{y} = H\mathbf{v} + \mathbf{z}, \tag{1.10}$$

where $\mathbf{z}$ is the error vector. The problem is now reduced to to: Fit $\mathbf{y}$ to be as close to $H\mathbf{v} + \mathbf{z}$ as possible, or to minimize the norm of the error vector, or

$$\begin{aligned} \min_{\mathbf{v}} \|\mathbf{y} - H\mathbf{v}\|_2^2 &= \min_{\mathbf{v}} (\mathbf{y} - H\mathbf{v})^T (\mathbf{y} - H\mathbf{v}) \\ &= \min_{\mathbf{v}} \left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T H\mathbf{v} - \mathbf{v}^T H\mathbf{y} + \mathbf{v}^T H^T H\mathbf{v} \right) \\ &= \min_{\mathbf{v}} \left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T H\mathbf{v} + \mathbf{v}^T H^T H\mathbf{v} \right). \end{aligned} \tag{1.11}$$

It is now possible to take the derivative with respect to the $\mathbf{v}$ vector (i.e. the gradient with respect to the coordinates of the constraint vector)

$$\begin{aligned} \frac{\partial}{\partial \mathbf{v}} \left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T H\mathbf{v} + \mathbf{v}^T H^T H\mathbf{v} \right) &= -2\mathbf{y}^T H + 2\mathbf{v}^T H^T H \\ &= 0, \end{aligned} \tag{1.12}$$

or

$$(H^T H)\mathbf{v} = H^T \mathbf{y}, \tag{1.13}$$

so, assuming that $H^T H$ is invertible, the optimization problem has solution

$$\mathbf{v}^* = (H^T H)^{-1} H^T \mathbf{y}, \tag{1.14}$$

where

$$\begin{aligned} H^T H &= \begin{bmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \\ &= \begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{bmatrix}, \end{aligned} \tag{1.15}$$

as seen in the calculus approach.

4

1.5 Maximum Likelyhood Estimation (MLE).

It is reasonable to ask why the 2-norm was picked for the objective function?

- One justification is practical: Because we can solve the derivative equation.

- Another justification: In statistics the error vector $\mathbf{z} = \mathbf{y} - H\mathbf{v}$ can be modelled as an IID (Independently and Identically Distributed) Gaussian random variable (i.e. noise). Under this model, the use of the 2-norm can be viewed as a consequence of such an ML estimation problem (see [1] ch. 7).

A Gaussian fig. 1.4 IID model is given by

$$y_i = ax_i + b \tag{1.16a}$$

$$z_i = y_i - ax_i - b \sim N(O, O^2) \tag{1.16b}$$

$$P_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}z^2/\sigma^2\right). \tag{1.16c}$$
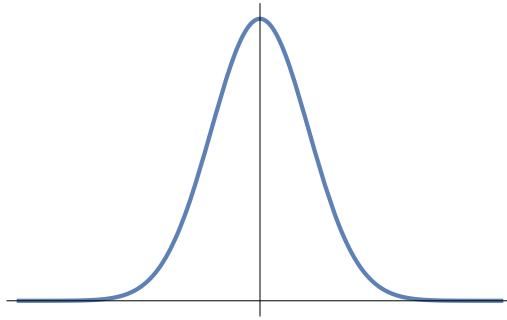

Figure 1.4: Gaussian probability distribution.

MLE: Maximum Likelyhood Estimator   Pick (a, b) to maximize the probability of observed data.

$$
\begin{aligned}
(a^*, b^*) &= \arg\max P(x, y; a, b) \\
&= \arg\max P_Z(y - (ax + b)) \\
&= \arg\max \prod_{i=1}^{n} \\
&= \arg\max \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(y_i - ax_i - b)^2/\sigma^2\right).
\end{aligned}
\tag{1.17}
$$

Taking logs gives

$$(a^*, b^*) = \arg\max \left(\text{constant} - \frac{1}{2} \sum_i (y_i - ax_i - b)^2/\sigma^2 \right)$$

$$= \arg\min \frac{1}{2} \sum_i (y_i - ax_i - b)^2/\sigma^2 \qquad (1.18)$$

$$= \arg\min \sum_i (y_i - ax_i - b)^2/\sigma^2$$

Here $\arg\max$ is not the maximum of the function, but the value of the parameter (the argument) that maximizes the function.

Double sides exponential noise   A double sided exponential distribution is plotted in fig. 1.5, and has the mathematical form

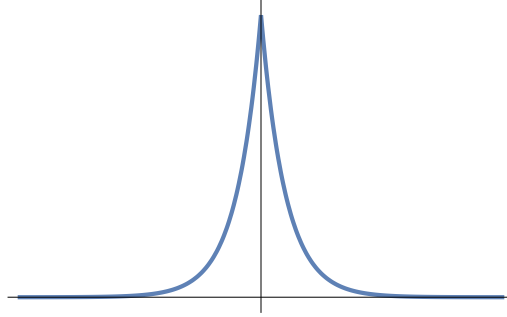$$P_Z(z) = \frac{1}{2c} \exp\left(-\frac{1}{c}|z| \right). \qquad (1.19)$$

Figure 1.5: Double sided exponential probability distribution.

The optimization problem is

$$\max_{a,b} \prod_{i=1}^{n} P_z(z_i) = \max_{a,b} \prod_{i=1}^{n} \frac{1}{2c} \exp\left(-\frac{1}{c}|z_i| \right)$$

$$= \max_{a,b} \prod_{i=1}^{n} \frac{1}{2c} \exp\left(-\frac{1}{c}|y_i - ax_i - b| \right) \qquad (1.20)$$

$$= \max_{a,b} \left(\frac{1}{2c} \right)^n \exp\left(-\frac{1}{c} \sum_{i=1}^{n} |y_i - ax_i - b| \right).$$

This is a L1 norm problem

$$\min_{a,b} \sum_{i=1}^{n} |y_i - ax_i - b|. \qquad (1.21)$$

i.e.

$$\min_{\mathbf{v}} \|\mathbf{y} - H\mathbf{v}\|_1. \tag{1.22}$$

This is still convex, but has no analytic solution, and is an example of a linear program.

1.5.1 *Solution of linear program*

Introduce helper variables $t_1, \cdots, t_n$, and minimize $\sum_i t_i$, such that

$$|y_i - ax_i - b| \leq t_i, \tag{1.23}$$

This is now an optimization problem for $a, b, t_1, \cdots t_n$. A linear program is defined as

$$\min_{a,b,t_1,\cdots t_n} \sum_i t_i \tag{1.24}$$

such that

$$y_i - ax_i - b \leq t_i y_i - ax_i - b \geq -t_i \tag{1.25}$$

Single sided exponential   What if your noise doesn't look double sided, with only noise for values $x > 0$. Can define a single sided probability distribution, as that of fig. 1.6.
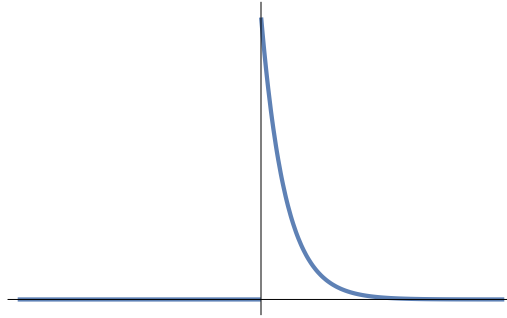

Figure 1.6: Single sided exponential distribution.

$$P_Z(z) = \begin{cases} \frac{1}{c}e^{-z/c} & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{1.26}$$

i.e. all z_i error values are always non-negative.

$$\log P_z(z) = \begin{cases} \text{const} - z/c & z > 0 \\ -\infty & z < 0 \end{cases} \tag{1.27}$$

Problem becomes

$$\min_{a,b} \sum_i \left(y_i - ax_i - b \right) \tag{1.28}$$

such that

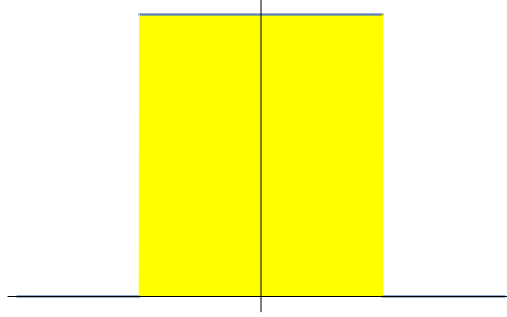$$y_i - ax_i - b \geq t_i \qquad \forall i \tag{1.29}$$

7

Figure 1.7: Uniform probability distribution.

Uniform noise   For noise that is uniformly distributed in a range, as that of fig. 1.7, which is constant in the range $[-c, c]$ and zero outside that range.

$$P_Z(z) = \begin{cases} \frac{1}{2c} & |z| \leq c \\ 0 & |z| > c. \end{cases} \tag{1.30}$$

or

$$\log P_Z(z) = \begin{cases} \text{const} & |z| \leq c \\ -\infty & |z| > c. \end{cases} \tag{1.31}$$

MLE solution

$$\max_{a,b} \prod_{i=1}^{n} P(x, y; a, b) = \max_{a,b} \sum_{i=1}^{n} \log P_Z(y_i - ax_i - b) \tag{1.32}$$

Here the argument is constant if $-c \leq y_i - ax_i - b \leq c$, so an ML solution is <u>any</u> (a, b) such that

$$|y_i - ax_i - b| \leq c \qquad \forall i \in 1, \cdots, n. \tag{1.33}$$

This is a linear program known as a "feasibility problem".

$$\min d \tag{1.34}$$

such that

$$\begin{aligned} y_i - ax_i - b &\leq d \\ y_i - ax_i - b &\geq -d \end{aligned} \tag{1.35}$$

If $d^* \leq c$, then the problem is feasible, however, if $d^* > c$ it is infeasible.

1.5.2   *Method comparison*

The double sided exponential, single sided exponential and uniform probability distributions of fig. 1.8 each respectively represent the point plots of the form fig. 1.9. The double sided exponential samples are distributed on both sides of the line, the single sided strictly above or on the line, and the uniform representing error bars distributed around the line of best fit.
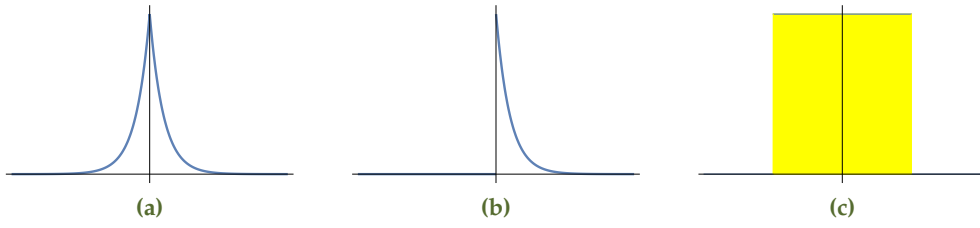
Figure 1.8: Distributions

Figure 1.9: Samples
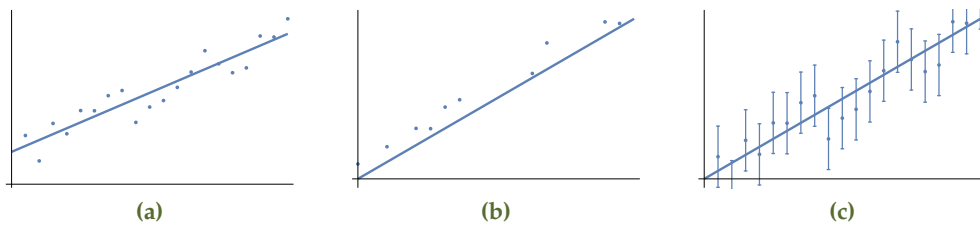
Bibliography

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
1.5