

32

Visualize Dirichlet Distribution

Dirichlet 分布

用投影法、重心坐标系可视化 Dirichlet 分布



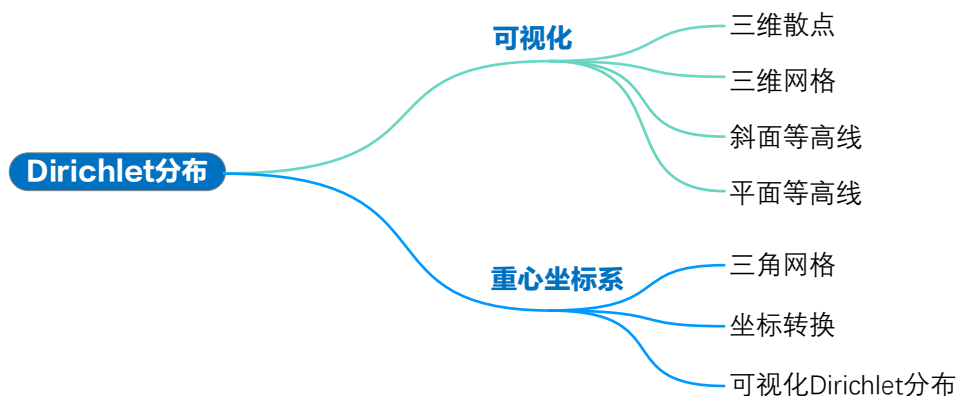
像专业人士一样学习规则，这样你就可以像艺术家一样打破它们。

Learn the rules like a pro, so you can break them like an artist.

—— 毕加索 (Pablo Picasso) | 西班牙艺术家 | 1881 ~ 1973



- ◀ matplotlib.pyplot.plot_trisurf() 在三角形网格上绘制平滑的三维曲面图
- ◀ matplotlib.pyplot.tricontourf() 在三角形网格上绘制填充的等高线图
- ◀ matplotlib.pyplot.triplot() 在三角形网格上绘制线条
- ◀ matplotlib.tri.Triangulation() 生成三角剖分对象
- ◀ matplotlib.tri.UniformTriRefiner() 对三角形网格进行均匀细化，生成更密集的三角形网格，以提高绘制的精细度和准确性
- ◀ numpy.column_stack() 将两个矩阵按列合并
- ◀ numpy.linalg.inv() 计算矩阵逆
- ◀ plotly.figure_factory.create_ternary_contour() 重心系等高线
- ◀ scipy.spatial.Delaunay() 生成一个点集的 Delaunay 三角剖分
- ◀ scipy.stats.dirichlet Dirichlet 分布对象
- ◀ scipy.stats.dirichlet.pdf() 计算 Dirichlet 分布的概率密度函数



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

32.1 什么是 Dirichlet 分布？

鸢尾花书第一册《数学要素》和本册《可视之美》前文中，大家最常见的分布应该是高斯分布。

本章则介绍一种很“美丽”的分布——Dirichlet 分布。

专门开辟一章特别讲解 Dirichlet 分布一方面是因为这个分布的确很美；此外，Dirichlet 分布常常用在贝叶斯统计学、自然语言处理中，重要程度不亚于高斯分布。

实际上，本书前文已经在不同章节聊过 Dirichlet 分布。如图 1 (a) 所示，这个三元 Dirichlet 变量有三个，即 θ_1 、 θ_2 、 θ_3 。

θ_1 、 θ_2 、 θ_3 取值范围都是 $[0, 1]$ 。非常重要的一点是， θ_1 、 θ_2 、 θ_3 满足 $\theta_1 + \theta_2 + \theta_3 = 1$ 。如图 1 (a) 所示，在三维空间来看， $\theta_1 + \theta_2 + \theta_3 = 1$ 且考虑给定取值范围 $[0, 1]$ ，形状为图中浅蓝色斜面。

给定三元 Dirichlet 的参数 $(\alpha_1, \alpha_2, \alpha_3)$ ，我们便可以获得三元 Dirichlet 的概率密度。如图 1 (b) 所示，本书前文用三维散点图展示过 Dirichlet 概率密度。图中每个散点的颜色代表概率密度大小，暖色更大，冷色更小。

还有一个有趣的知识点是，为了方便可视化 Dirichlet 分布，我们可以顺便了解重心坐标系这个概念，这是本书最后要介绍的内容。

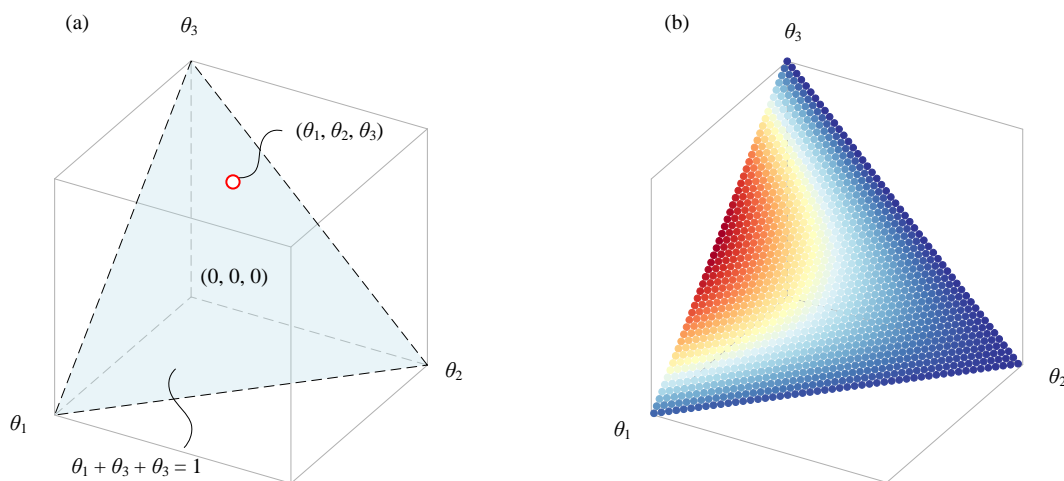


图 1. 用三维展示 Dirichlet 分布

32.2 降维投影到平面

由于 $\theta_1 + \theta_2 + \theta_3 = 1$ 这个约束条件，任意给定两个 θ ，我们就可以确定斜面上一点。

如图 2 所示， $\theta_1 + \theta_2 + \theta_3 = 1$ 代表浅蓝色斜面。当给定 θ_1 和 θ_2 时，我们便可以计算得到 θ_3 。也就是说，我们可以将浅蓝色斜面投影到 $\theta_1\theta_2$ 平面上， $(\theta_1, \theta_2, \theta_3)$ 和 (θ_1, θ_2) 一一对应。

显然，在 $\theta_1\theta_2$ 平面上绘制等高线很容易！

图 9 所示为一组在平面上展示的 Dirichlet 分布。Bk_2_Ch32_01.ipynb 绘制这组图，代码很简单，请大家自行学习。

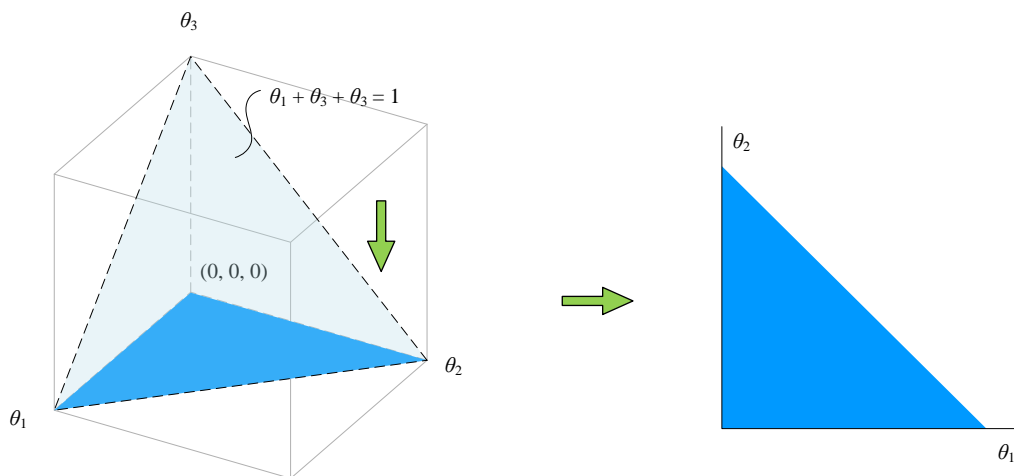


图 2. 浅蓝色平面 $\theta_1 + \theta_2 + \theta_3 = 1$ 投影到 $\theta_1\theta_2$ 平面

32.3 将等高线投影到斜面上

由于 $(\theta_1, \theta_2, \theta_3)$ 和 (θ_1, θ_2) 一一对应，我们立刻就想到，我们是否能把图 9 这些平面等高线上投影到 $\theta_1 + \theta_2 + \theta_3 = 1$ 上？

答案是肯定的！

图 2 所示的就是这个意义映射原理。

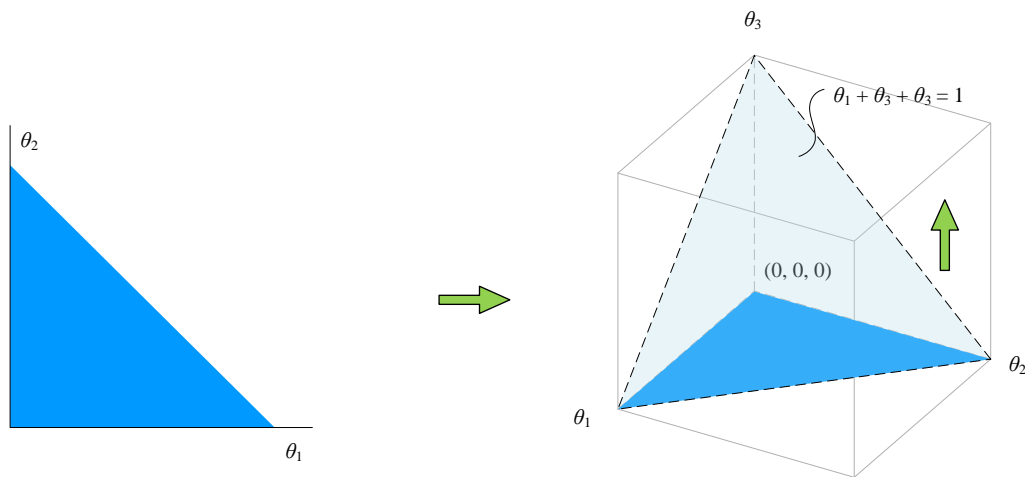


图 3. $\theta_1\theta_2$ 平面反过来投影到浅蓝色平面 $\theta_1 + \theta_2 + \theta_3 = 1$

如图 10 所示，我们把图 9 中 Dirichlet 概率密度函数等高线投影到斜面上。这里用的的可视化技巧类似于前文在介绍瑞利商时，将瑞利商投影到单位球体“幕布”上几乎完全一致。

Bk_2_Ch32_02.ipynb 绘制图 10，核心代码已经在本书前文将结果，也请大家自行学习。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

类似地，如图 4 所示，我们可以设计更丰富的可视化方案用不同投影方法展示 Dirichlet 分布。图 11、图 12、图 13 这三组图就是可视化的结果。用到的绘图技巧是 Matplotlib 中三维等高线在不同方向投影，本书在前文介绍过这个技巧，大家可以回顾。

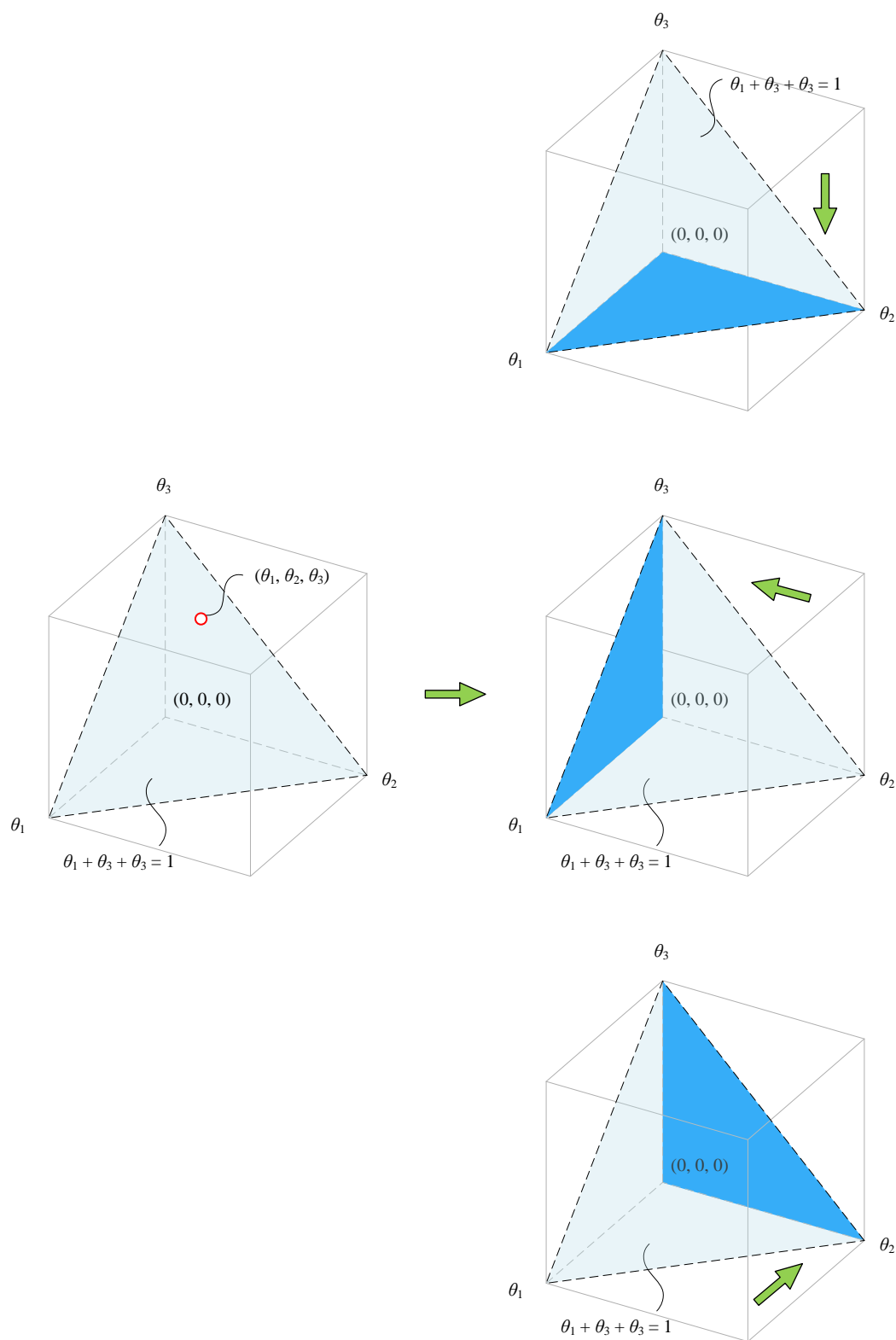


图 4. 投影到三个不同平面上

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

32.4 重心坐标系

从物理角度来看，**重心坐标系** (barycentric coordinate system) 是一种描述一个几何形状内部任意点位置的方法。它是将该形状的重心作为原点建立的坐标系。

在平面上的一个三角形中，任何一点都可以表示为三个定点的加权平均值，其中每个定点的权重由它到该点的距离与该三角形的周长之比确定。这些权重称为该点在三角形的重心坐标。

实际上，用三维直角坐标系解释重心坐标系更方便。图 5 左图所示为三维直角坐标系，为了区分坐标系的横轴、纵轴、竖轴分别记做 θ_1 、 θ_2 、 θ_3 。这个三维直角坐标系中坐标可以记做 $(\theta_1, \theta_2, \theta_3)$ 。

图 5 左图浅蓝色平面上的每个坐标 $(\theta_1, \theta_2, \theta_3)$ 都满足 $\theta_1 + \theta_2 + \theta_3 = 1$ 。这个限制条件，让原本三维的空间降维成二维。即便如此，如图 5 右图所示，三角网格的每一点仍旧对应 $(\theta_1, \theta_2, \theta_3)$ 。

如图 6 所示，本章并用两种变量 θ_1 、 θ_2 、 θ_3 标注位置。

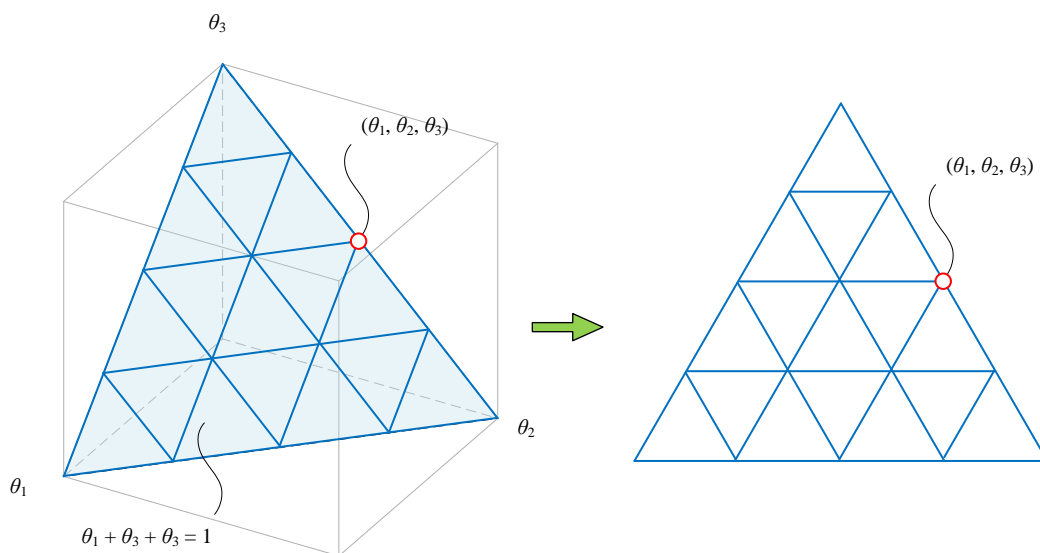


图 5. 从三维直角坐标系到重心坐标系

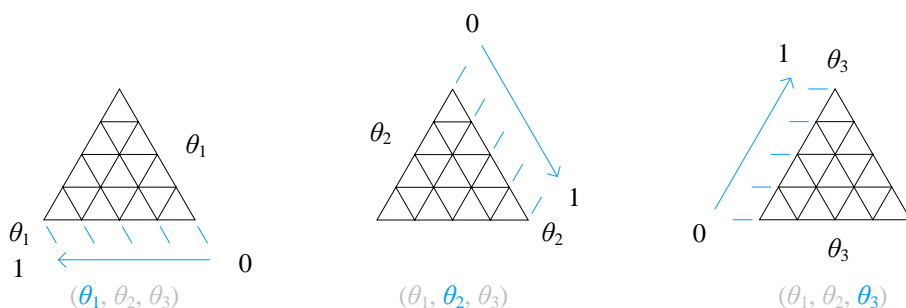


图 6. 重心坐标系的三个变量

图 14 从三维直角坐标系视角观察中心坐标系中的 θ_1 、 θ_2 、 θ_3 。

图 15 上下两幅子图比较**等边三角形** (equilateral triangle) 中的两个坐标系坐标关系；上图给出的是利用三角网格表达平面直角坐标系。下图则是利用相同网格表达重心坐标系坐标。请大家注意区分。

图 16 则展示**等腰直角三角形** (isosceles right triangle) 中两个坐标系坐标关系。图 17 所示为任意形状三角形中两个坐标系的关系。

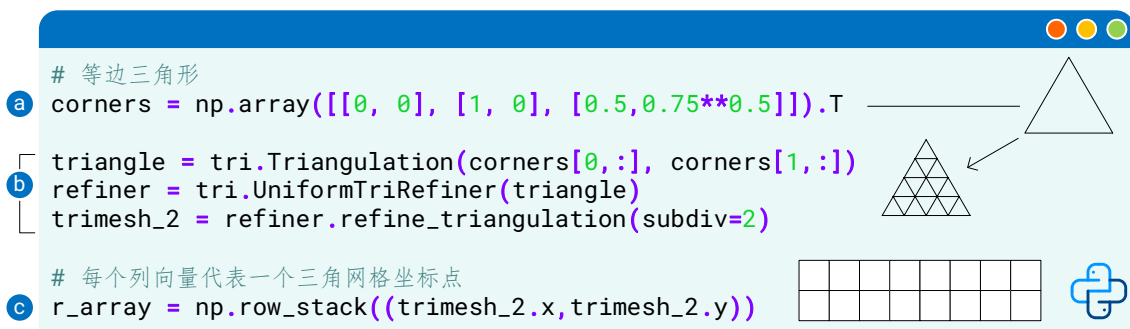
Bk_2_Ch32_06.ipynb 绘制图 15、图 16、图 17，下面聊聊其中核心语句。

代码 1 在直角坐标系中生成三角网格坐标点。有了本章前文的基础，大家应该对这几句代码很熟了。

a 定义三角形的三个顶点坐标，这个三角形是图 15 中等边三角形。Bk_2_Ch32_06.ipynb 还给出两个其他三角形的顶点坐标。

b 构造三角网格对象，然后细分网格。

c 将三角网格坐标点构造成二维数组（矩阵），每个列向量代表一个坐标点。



```
# 等边三角形
a corners = np.array([[0, 0], [1, 0], [0.5, 0.75**0.5]]).T

b triangle = tri.Triangulation(corners[0,:], corners[1,:])
  refiner = tri.UniformTriRefiner(triangle)
  trimesh_2 = refiner.refine_triangulation(subdiv=2)

# 每个列向量代表一个三角网格坐标点
c r_array = np.row_stack((trimesh_2.x, trimesh_2.y))
```

代码 1. 直角坐标系中的三角网格坐标点 | Bk_2_Ch32_06.ipynb

代码 2 可视化直角坐标系中网格坐标。

a 用 `matplotlib.pyplot.triplot()`，简作 `plt.triplot()`，绘制三角网格。

b 使用 `for` 循环在图片上打印三角网格每个点的坐标。

其中，`zip()` 将多个可迭代对象打包，从而方便 `for` 循环遍历。

每次迭代，创建 `text_idx`，其中包含横纵坐标的字符串。`format()` 将坐标值格式化为小数点后两位。

然后利用 `matplotlib.pyplot.text()`，简作 `plt.text()`，将文本添加到图片中。

`x_idx` 是文本标签的 `x` 坐标。

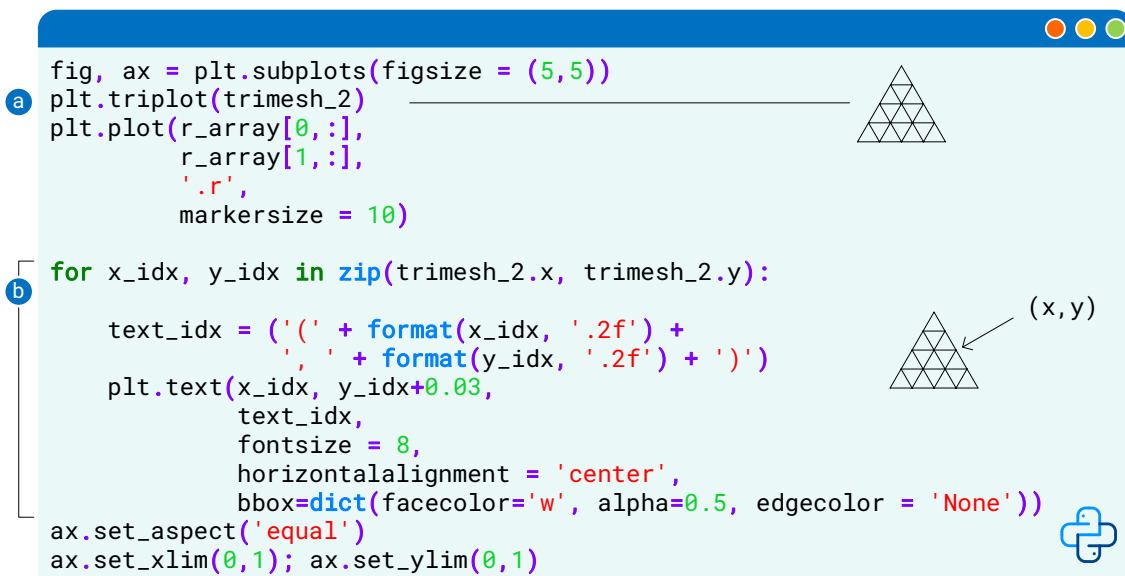
`y_idx+0.03` 是为了在 `y` 方向上将文本标签稍微上移，以免与数据点重叠。

`text_idx` 是要显示的文本字符串。

`fontsize=8` 设置文本的字体大小为 8 pt。

`horizontalalignment='center'` 将文本水平居中对齐。

`bbox=dict(facecolor='w', alpha=0.5, edgecolor='None')` 定义了一个包含文本的矩形框。其中 `facecolor='w'` 设置矩形框的背景色为白色, `alpha=0.5` 设置透明度为 0.5, `edgecolor='None'` 表示矩形框没有边框。



```
fig, ax = plt.subplots(figsize = (5,5))
a plt.triplot(trimesh_2)
plt.plot(r_array[0,:],
         r_array[1,:],
         '.r',
         markersize = 10)

b for x_idx, y_idx in zip(trimesh_2.x, trimesh_2.y):
    text_idx = '(' + format(x_idx, '.2f') +
               ', ' + format(y_idx, '.2f') + ')'
    plt.text(x_idx, y_idx+0.03,
             text_idx,
             fontsize = 8,
             horizontalalignment = 'center',
             bbox=dict(facecolor='w', alpha=0.5, edgecolor = 'None'))
ax.set_aspect('equal')
ax.set_xlim(0,1); ax.set_ylim(0,1)
```

代码 2. 展示直角坐标系中网格坐标 | Bk_2_Ch32_06.ipynb

代码 3 展示的是直角坐标系和重心坐标系坐标转换运算。图 7 所示为转换过程用到的具体数学工具, 大家可以看到其中最重要的运算是矩阵求逆。Bk_2_Ch32_06.ipynb 列出参考文献, 大家可以自行学习。

a 提取大三角形三个顶点坐标, 结果为二维数组, 相当于列向量。

b 首先创建矩阵 $T = [r_1 - r_3 \quad r_2 - r_3]$ 。

然后, 通过 $\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = T^{-1} (r - r_3)$ 计算 θ_1 、 θ_2 。代码中用到了广播原则。`r_array` 是直角坐标系中有待转换的坐标点。

根据等式 $\theta_1 + \theta_2 + \theta_3 = 1$ 计算 θ_3 。注意, `r_array` 和 `theta_1_2_3` 每一列一一对应。也就是说, 至此, 我们完成了直角坐标系和重心坐标系之间的坐标转换。

c 利用 `numpy.row_stack()` 创建一个矩阵, 每一列代表一个重心坐标坐标。

d 利用 `numpy.clip()` 对计算得到的重心坐标进行截断, 确保它们在合理的范围内, 避免超出 $[0, 1]$ 区间。`1e-6` 是一个小的容差值, 用于避免由于数值误差导致的坐标超出范围。

```

# 提取大三角形的三个顶点列向量（坐标点）
a r1 = corners[:, [0]]
  r2 = corners[:, [1]]
  r3 = corners[:, [2]]
# 构造矩阵T
b T = np.column_stack((r1 - r3, r2 - r3))
# 计算 theta_1 和 theta_2
  theta_1_2 = np.linalg.inv(T) @ (r_array - r3)
# 计算 theta_3
  theta_3 = 1 - theta_1_2[0, :] - theta_1_2[1, :]

# 创建 theta 坐标，每一列代表一个重心坐标系坐标
# r_array 和 theta_1_2_3 每一列一一对应
c theta_1_2_3 = np.row_stack((theta_1_2, theta_3))

# 对重心坐标进行截断，避免超出 [0, 1] 区间
d theta_1_2_3 = np.clip(theta_1_2_3, 1e-6, 1.0 - 1e-6)

```

代码 3. 直角坐标系和重心坐标系转换 | Bk_2_Ch32_06.ipynb

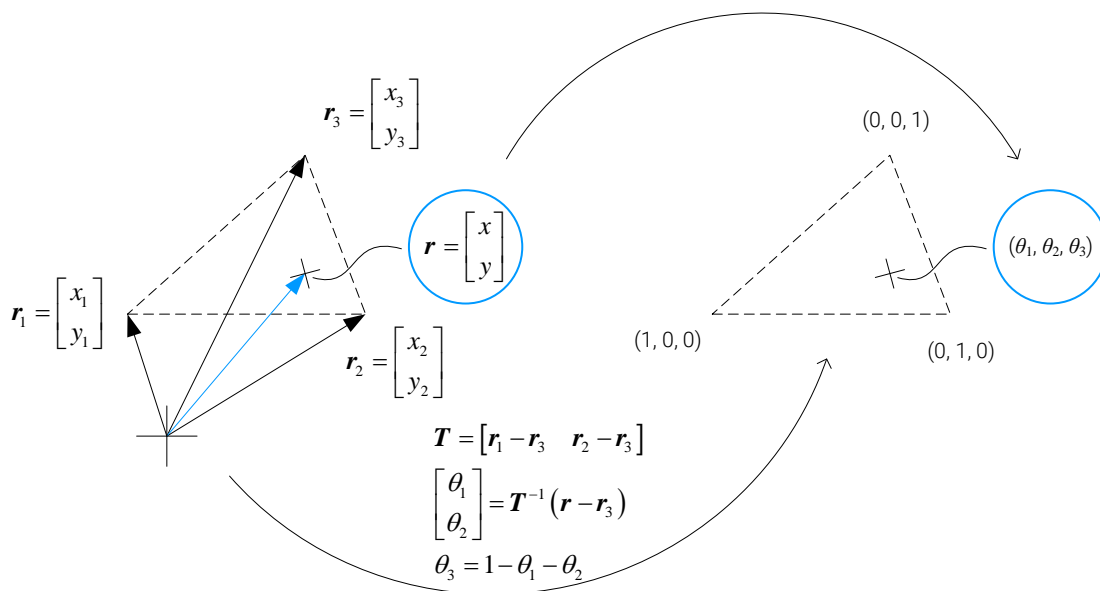


图 7. 平面直角坐标系到重心坐标系转换背后的线性代数工具

? 请大家修改 Bk_2_Ch32_06.ipynb，修改原始大三角形的三个顶点坐标为 $(0, 8)$ 、 $(8, 0)$ 、 $(5, 8)$ ，重新绘制重心坐标系网格坐标。

图 18 所示为在重心坐标系中混合红绿蓝三色。Bk_2_Ch32_07.ipynb 绘制图 18，代码很简单，请大家自行学习。

32.4 重心坐标系展示 Dirichlet 分布

“鸢尾花书”中，重心坐标系常用来可视化 Dirichlet 分布概率密度函数。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



《统计至简》一册将专门讲解 Dirichlet 分布及其在贝叶斯推断的应用。

Dirichlet 分布是一种连续的概率分布，通常用于描述一个多元随机变量的概率分布。它是以德国数学家 Peter Gustav Lejeune Dirichlet 的名字命名的。

本书前文利用其它可视化方案展示过 Dirichlet 分布。图 19、图 20 向大家展示如何利用重心坐标系可视化三元 Dirichlet 分布随参数 $[\alpha_1, \alpha_2, \alpha_3]$ 变化。

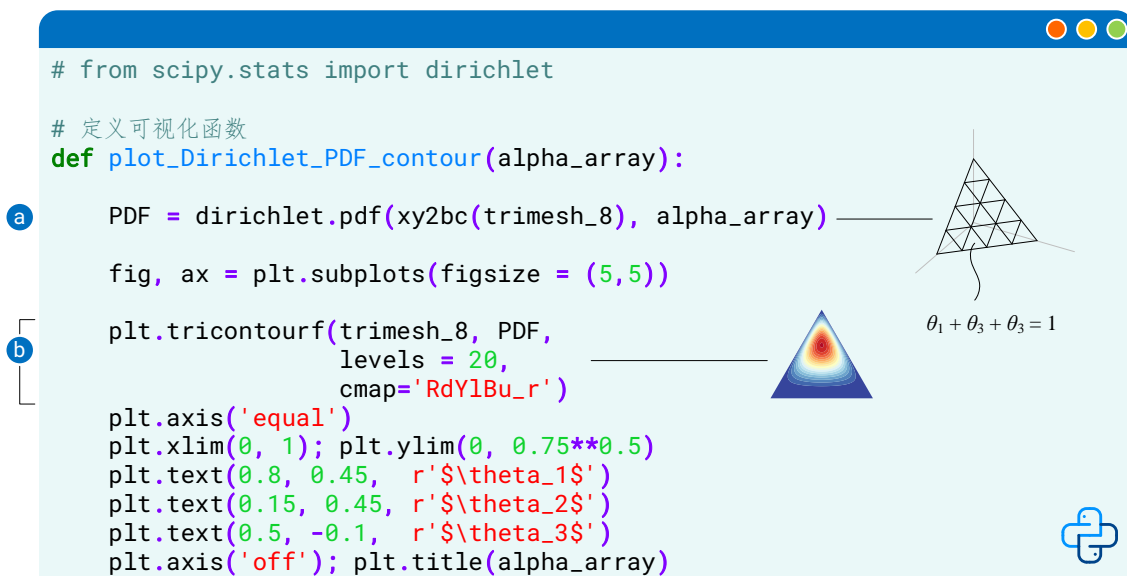
Bk_2_Ch32_08.ipynb 绘制图 19、图 20，下面聊聊其中关键语句。

代码 4 所示为自定义函数用来可视化 Dirichlet 分布。

a 用 `scipy.stats.dirichlet.pdf()`，简作 `dirichlet.pdf()`，计算给定重心坐标系网格上点的 Dirichlet 分布的概率密度函数值。这个函数的输入为网格坐标 $(\theta_1, \theta_2, \theta_3)$ ，分布参数 $(\alpha_1, \alpha_2, \alpha_3)$ 。其中， $\theta_1, \theta_2, \theta_3$ 非负，且满足 $\theta_1 + \theta_2 + \theta_3 = 1$ ，正是这个条件使得我们可以用重心坐标系可视化 Dirichlet 分布 PDF。

其中，`xy2bc()` 为自定义函数，将直角坐标系坐标转化为重心坐标系坐标。这个自定义函数的核心代码来自代码 3。

b 利用 `matplotlib.pyplot.tricontourf()`，简作 `plt.tricontourf()`，绘制基于三角形网格的等高线填充图。



代码 4. 可视化 Dirichlet 分布 | Bk_2_Ch32_08.ipynb

Plotly 中也有绘制基于三角形网格等高线的函数，具体图 21 所示。

对应的代码文件为 Bk_2_Ch32_09.ipynb。

利用的函数为 `plotly.figure_factory.create_ternary_contour()`；代码相对简单，请大家自行学习使用这个函数。

如图 8 所示，在 Bk_2_Ch32_09.ipynb 上，我们用 Streamlit 制作了一个 App，用来展示 Dirichlet 分布参数对概率密度函数图像影响。

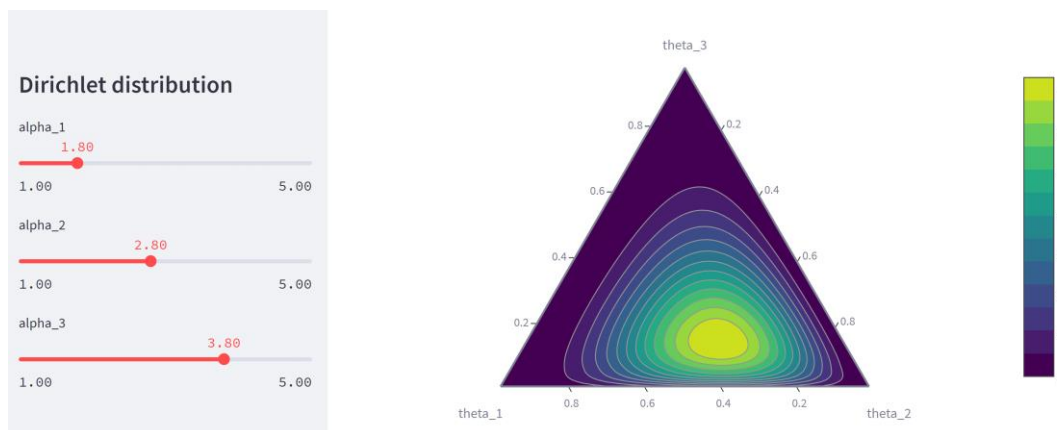
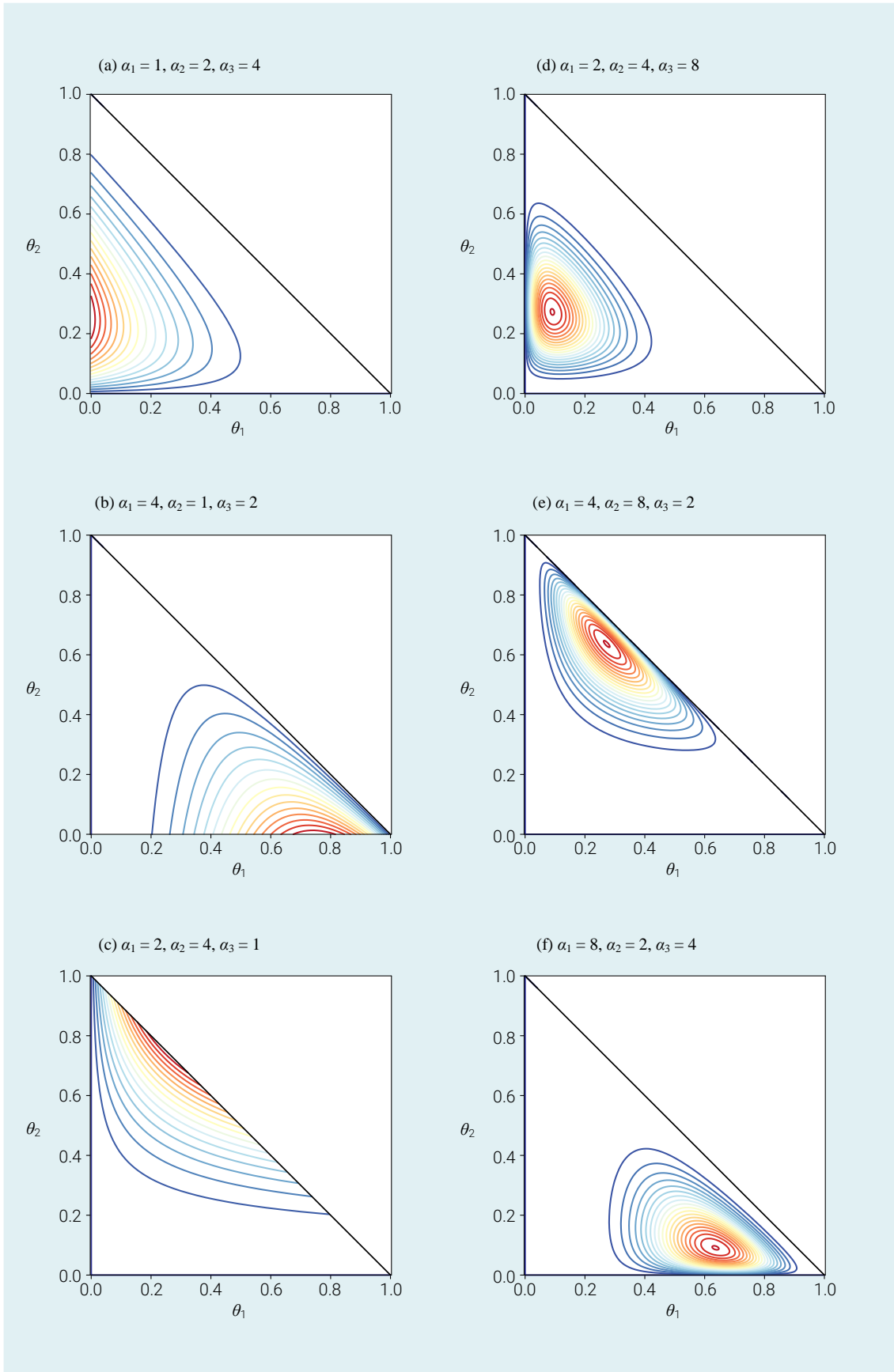


图 8. 展示 Dirichlet 分布的 App, Streamlit 搭建 | [Streamlit_Dirichlet_distribution.py](#)

本章设计了几种方案可视化 Dirichlet 分布。稍有挑战性的知识点是如何理解重心坐标系，以及直角坐标系和重心坐标系坐标的转换。

图 9. Dirichlet 分布, $\theta_1\theta_2$ 平面 |  Bk_2_Ch32_01.py

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

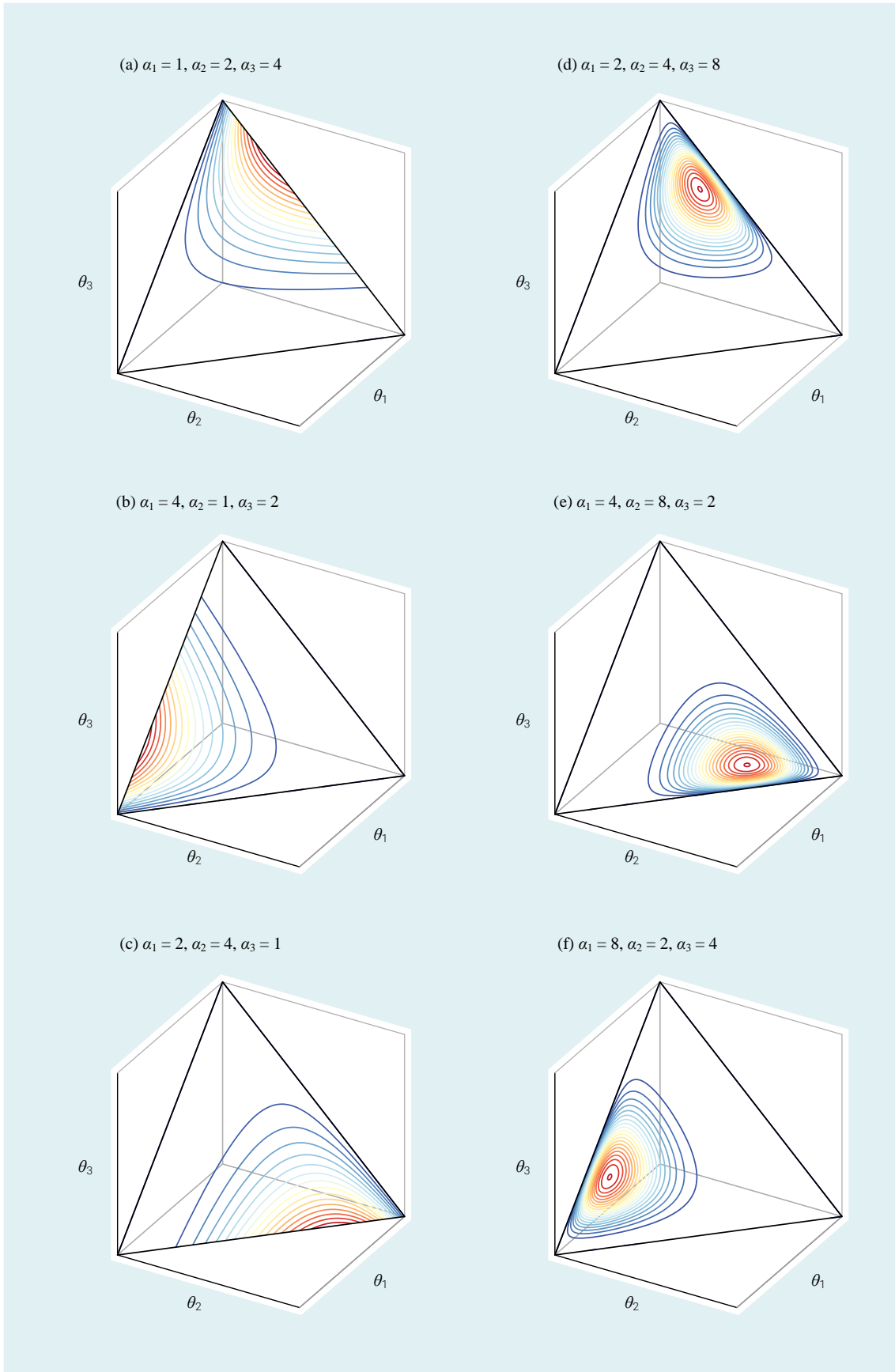



图 10. Dirichlet 分布，投影到斜面 |  Bk_2_Ch32_02.py

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

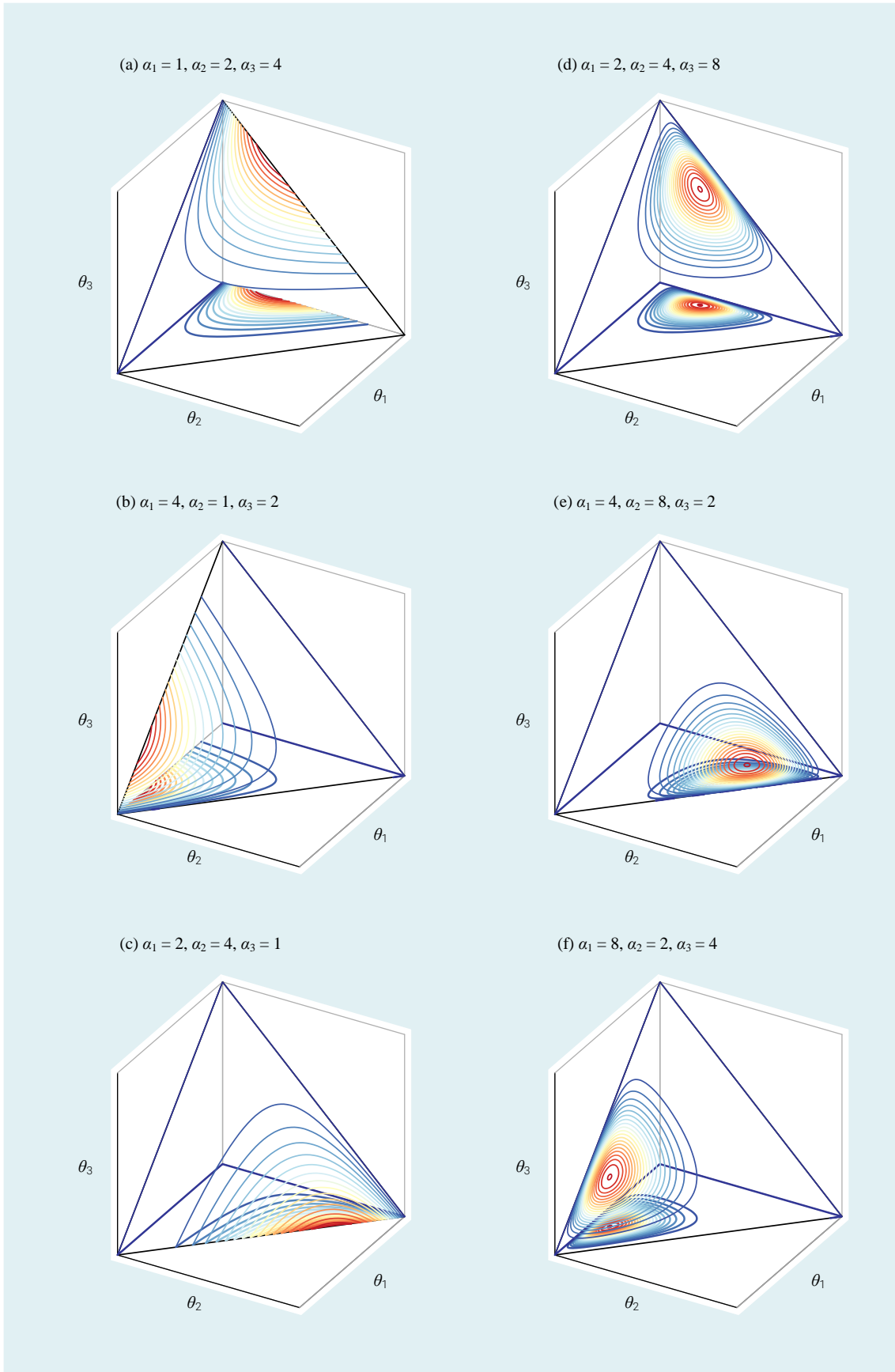



图 11. Dirichlet 分布, 斜面 + $\theta_1\theta_2$ 平面 |  Bk_2_Ch32_03.py

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

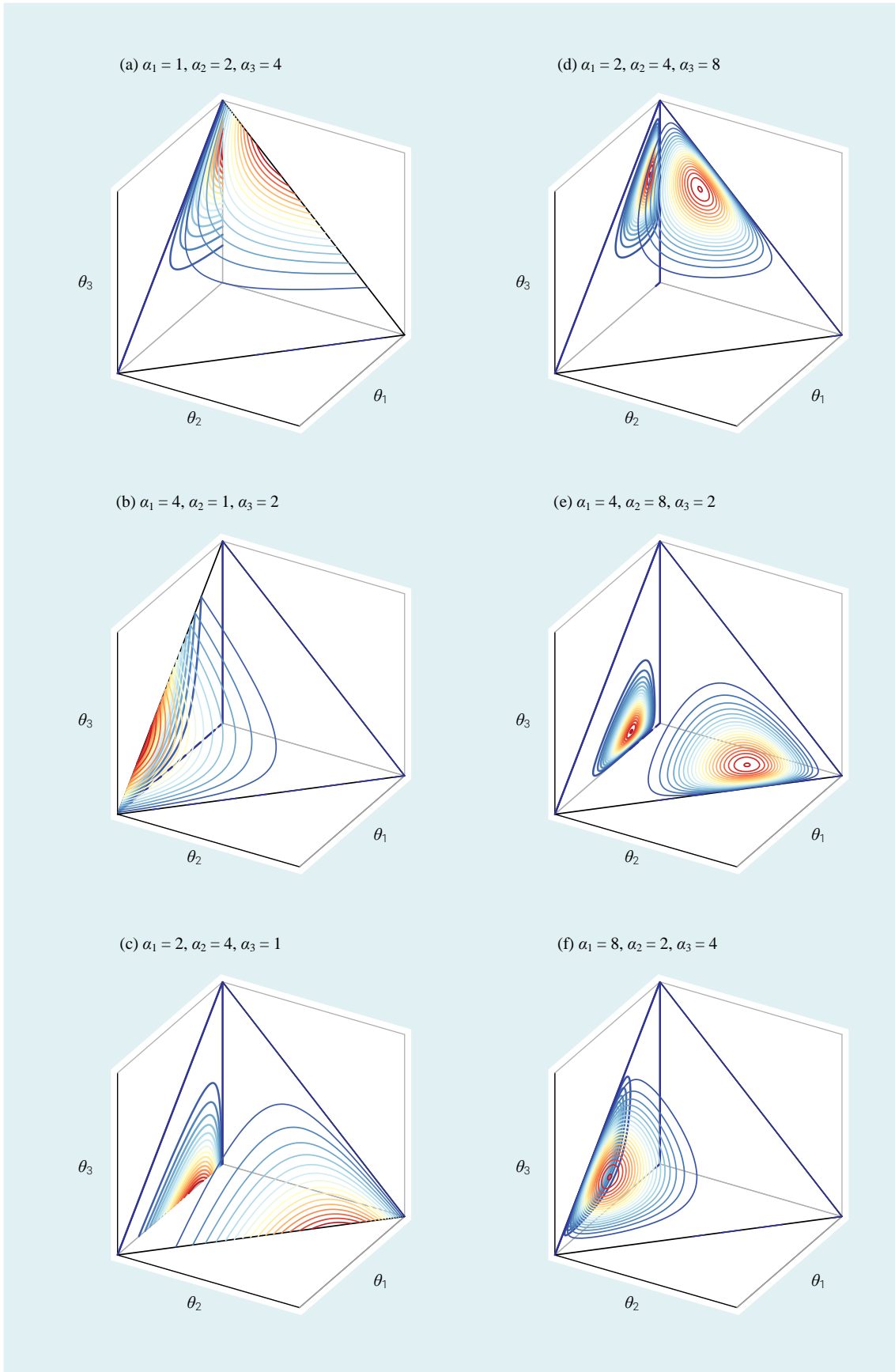



图 12. Dirichlet 分布, 斜面 + $\theta_1\theta_3$ 平面 |  Bk_2_Ch32_04.py

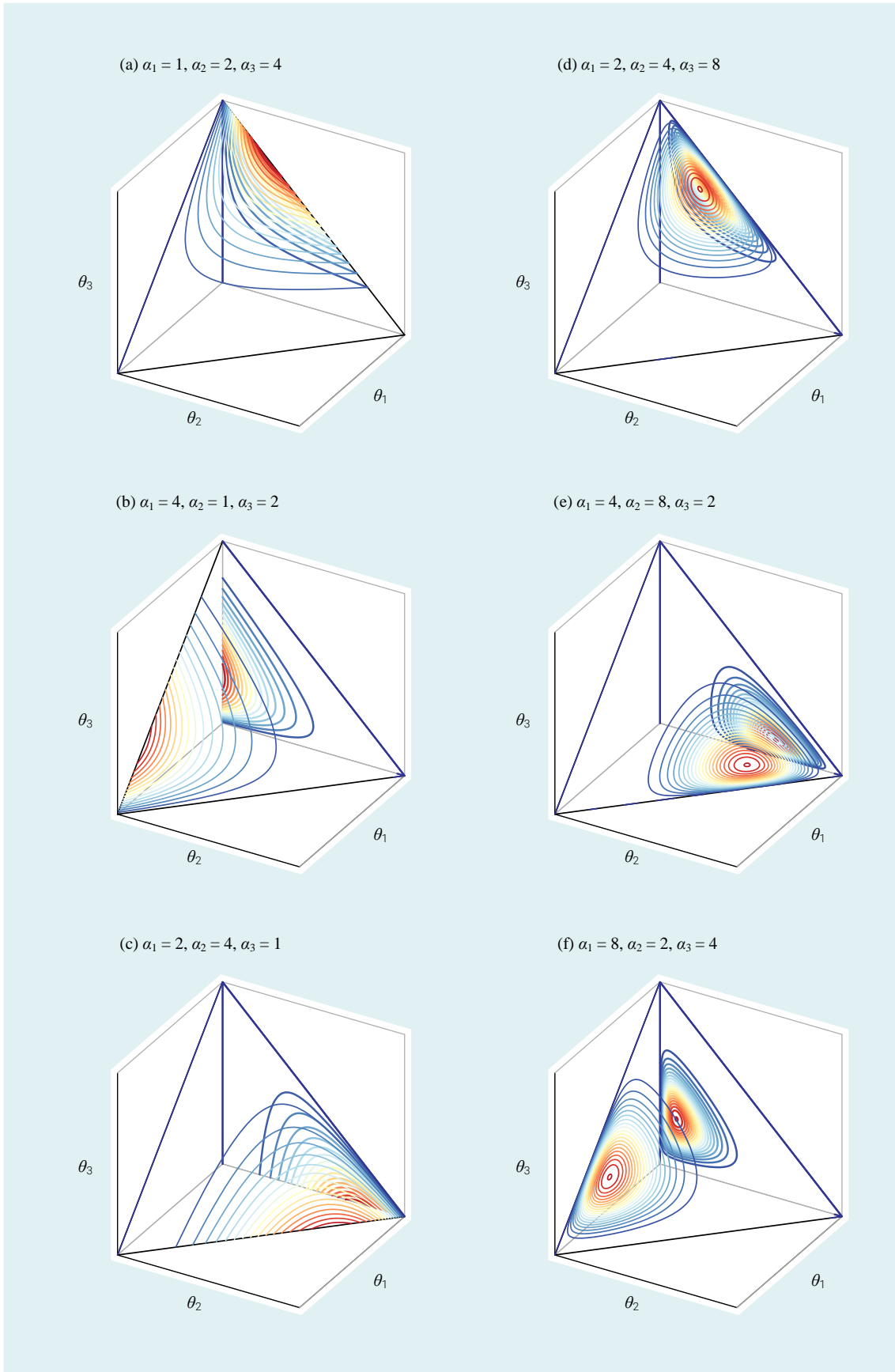



图 13. Dirichlet 分布, 斜面 + $\theta_2\theta_3$ 平面 |  Bk_2_Ch32_05.py

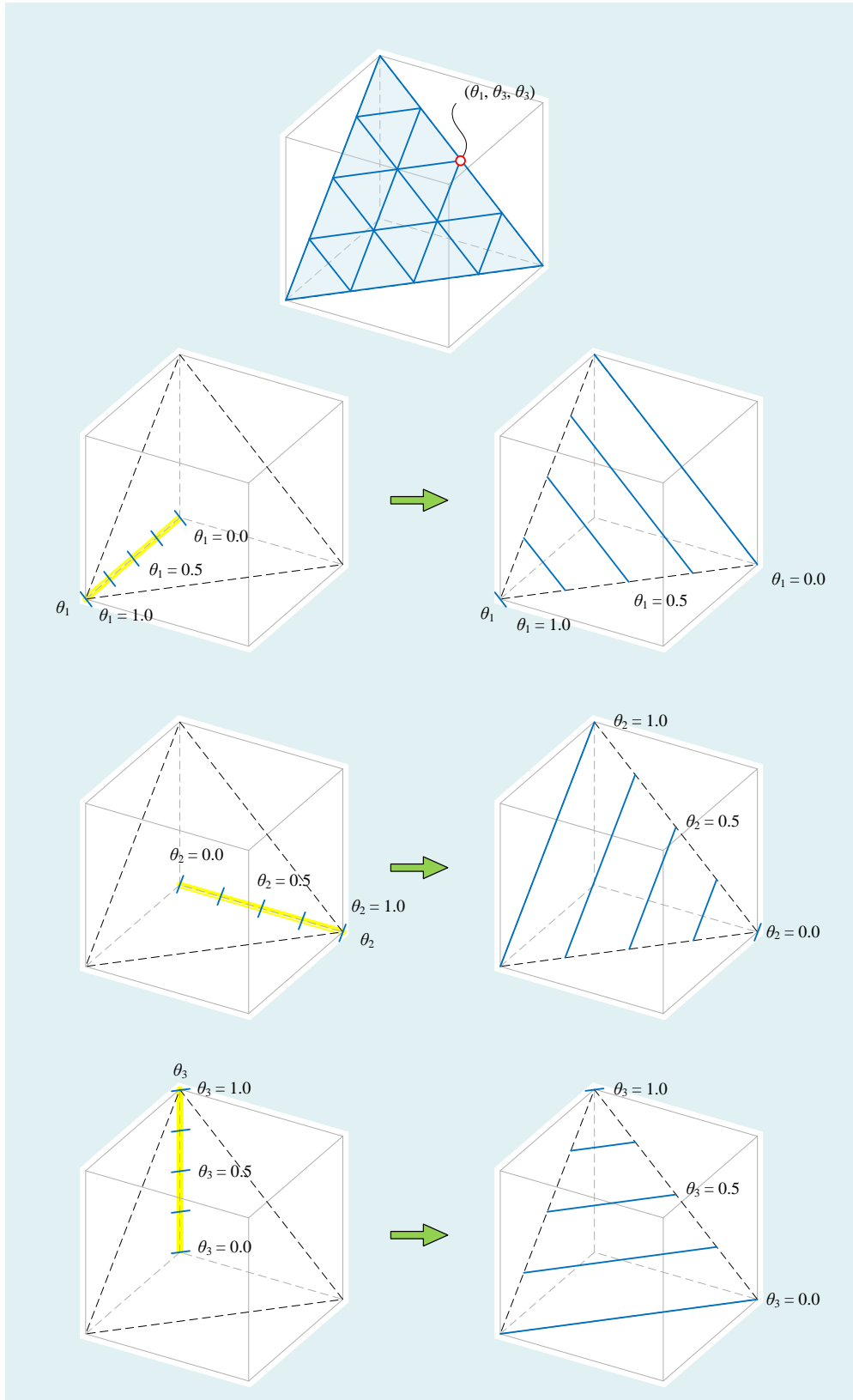


图 14. 三维直角坐标系角度看重心坐标系

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

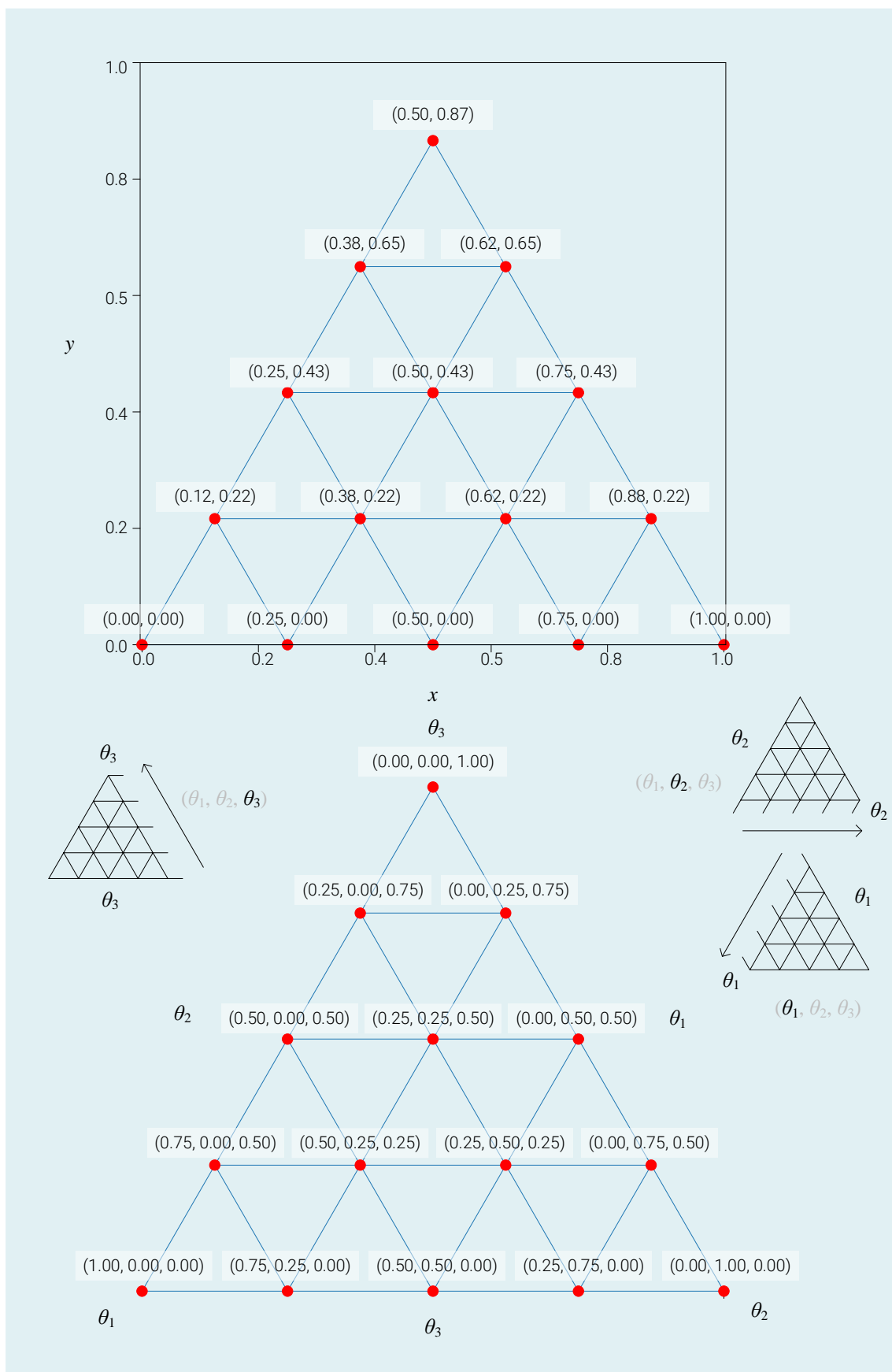


图 15. 比较平面直角坐标系坐标、重心坐标系坐标，等边三角形 | Bk_2_Ch32_06.ipynb

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

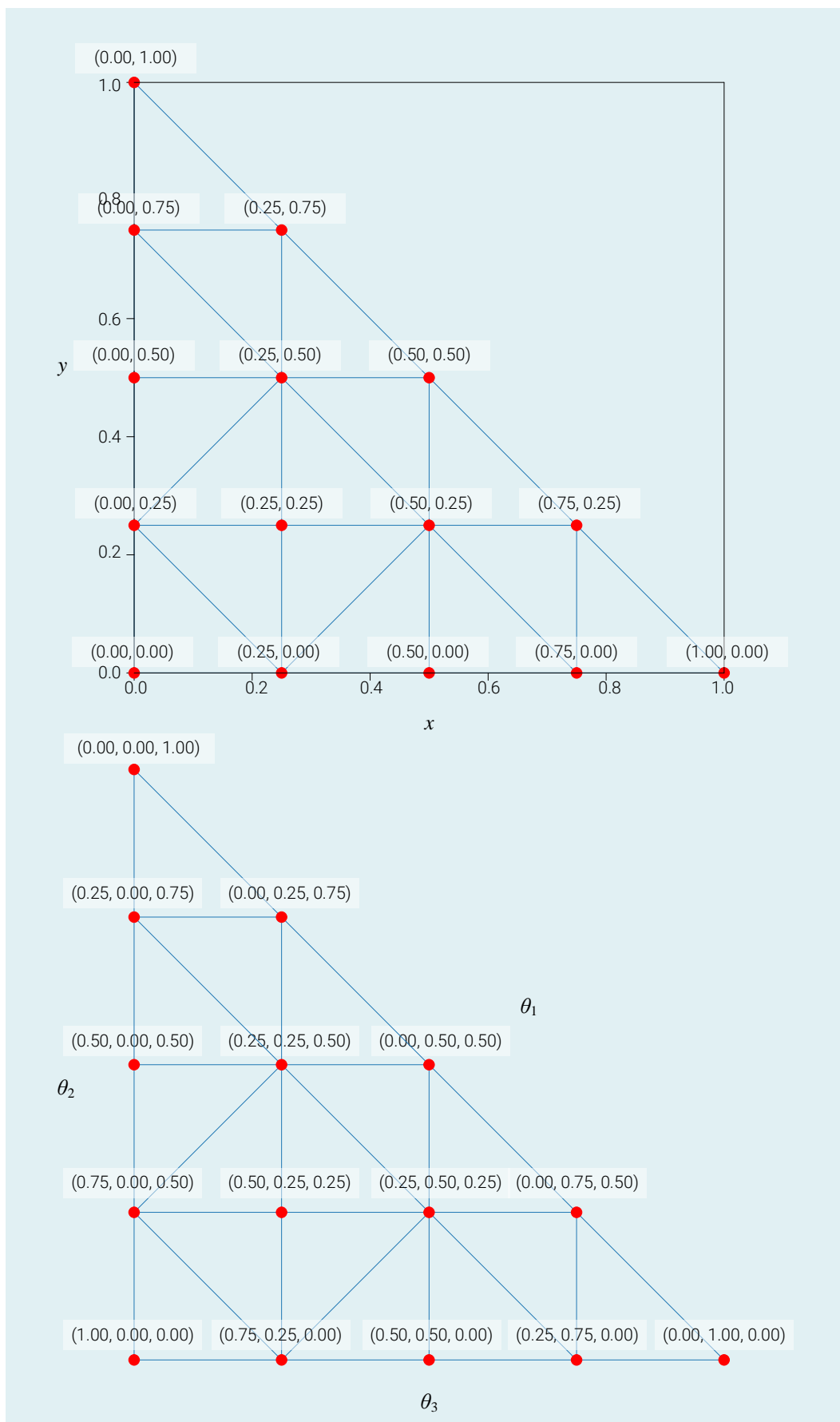


图 16. 比较平面直角坐标系坐标、重心坐标系坐标，等腰直角三角形 | Bk_2_Ch32_06.ipynb

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

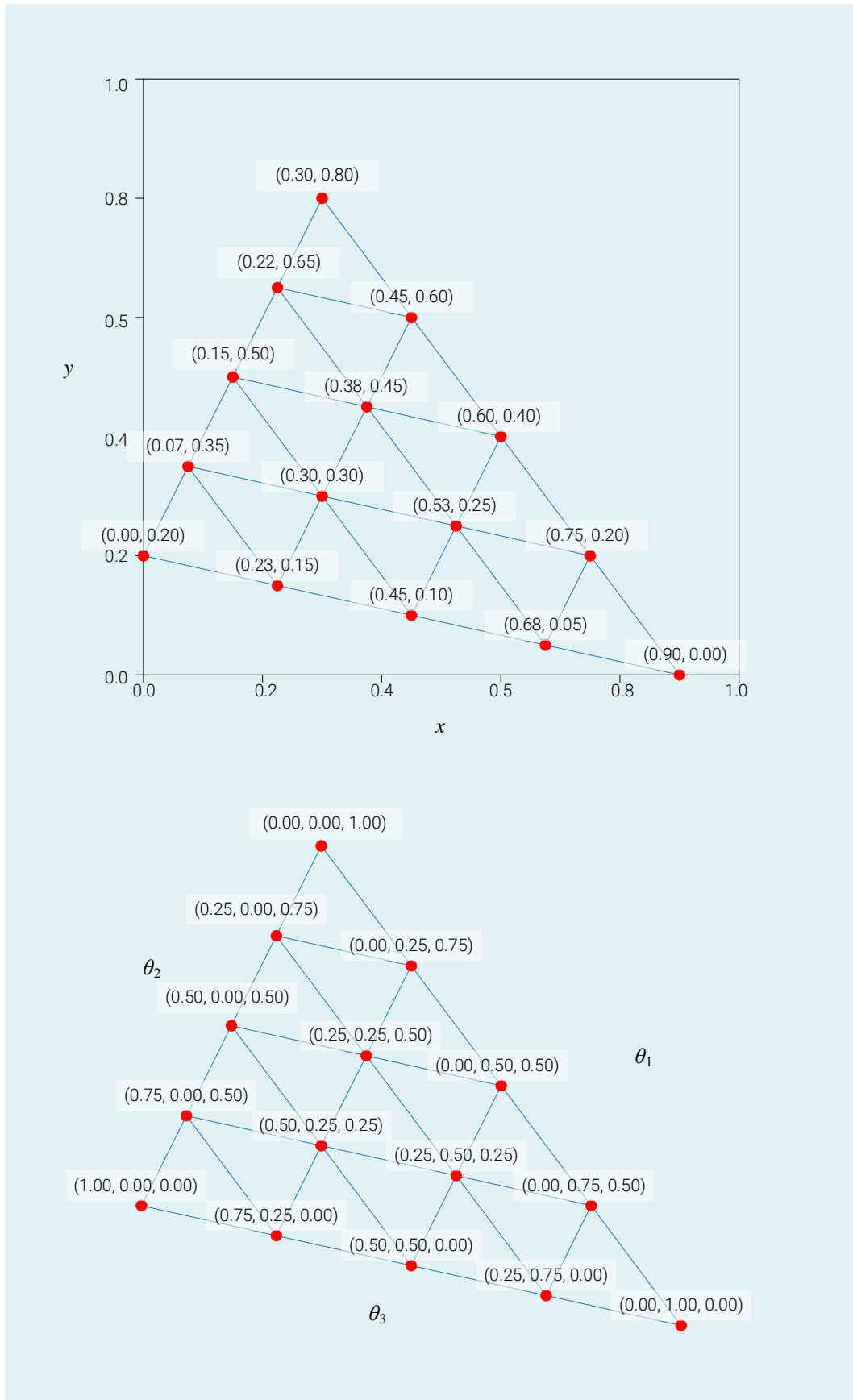


图 17. 比较平面直角坐标系坐标、重心坐标系坐标，任意形状三角形 | Bk_2_Ch32_06.ipynb

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

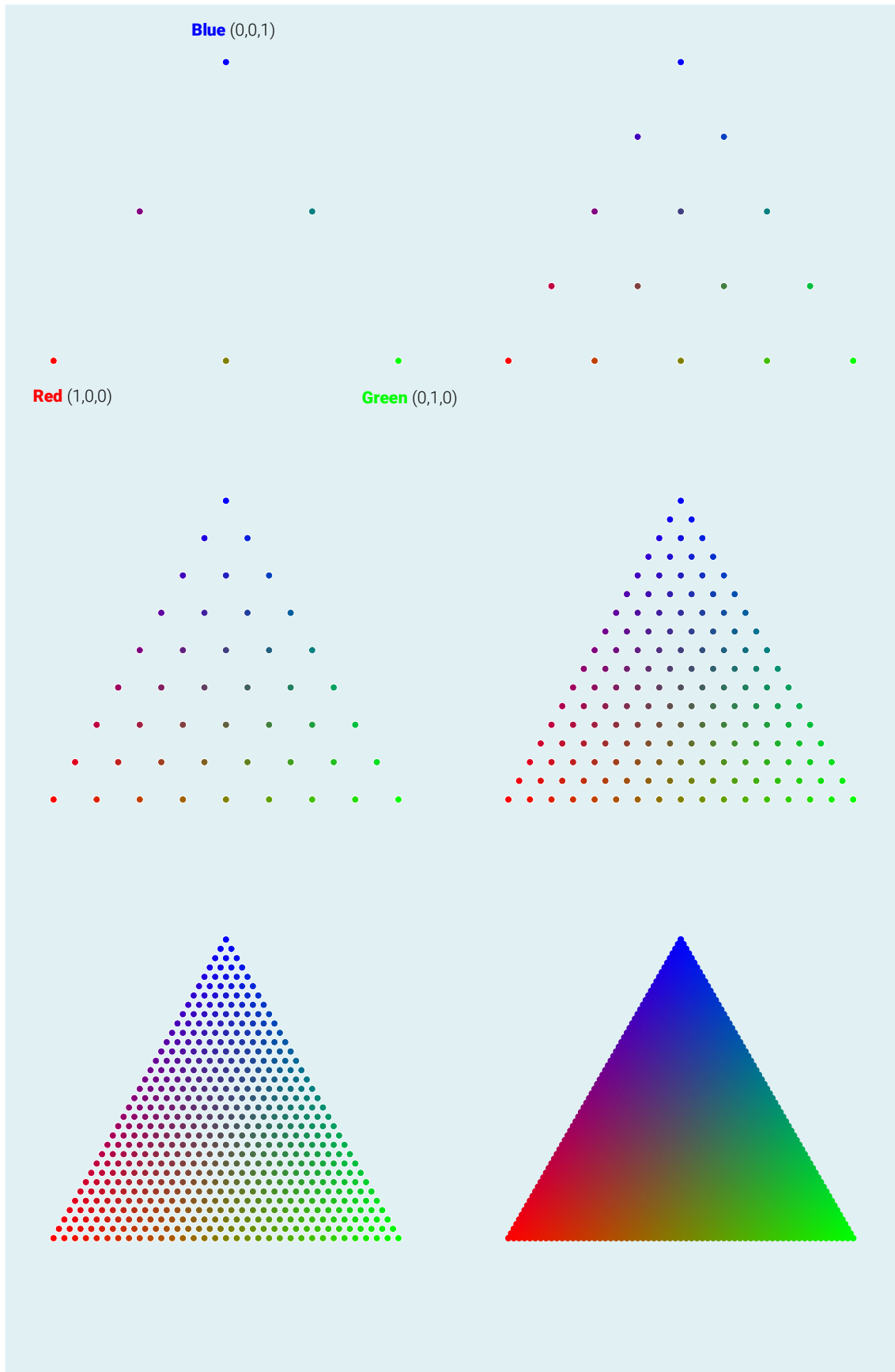


图 18. 重心坐标系中混合红绿蓝 | Bk_2_Ch32_07.ipynb

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

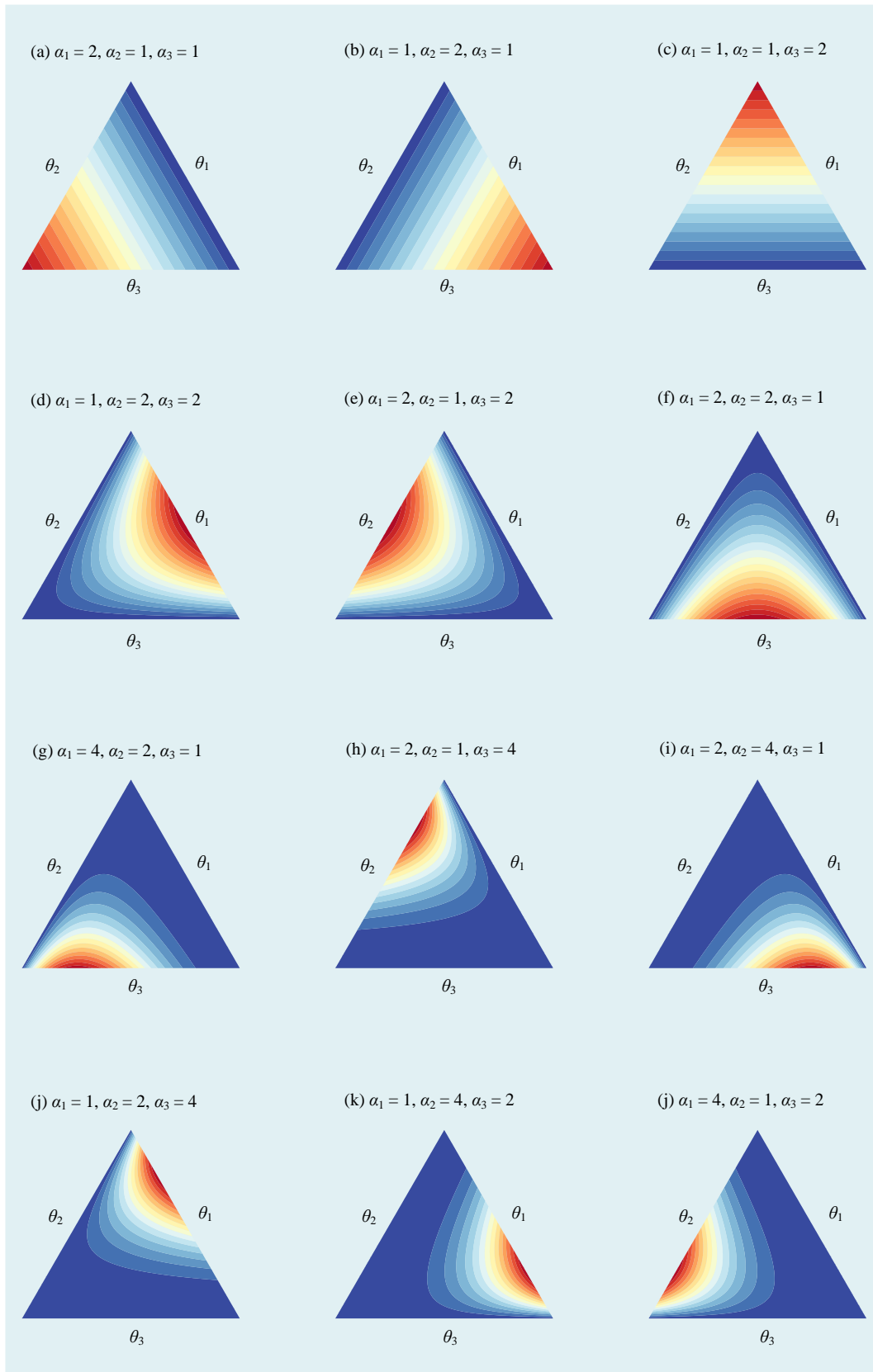


图 19. Dirichlet 分布, 第 1 组 | Bk_2_Ch32_08.ipynb

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

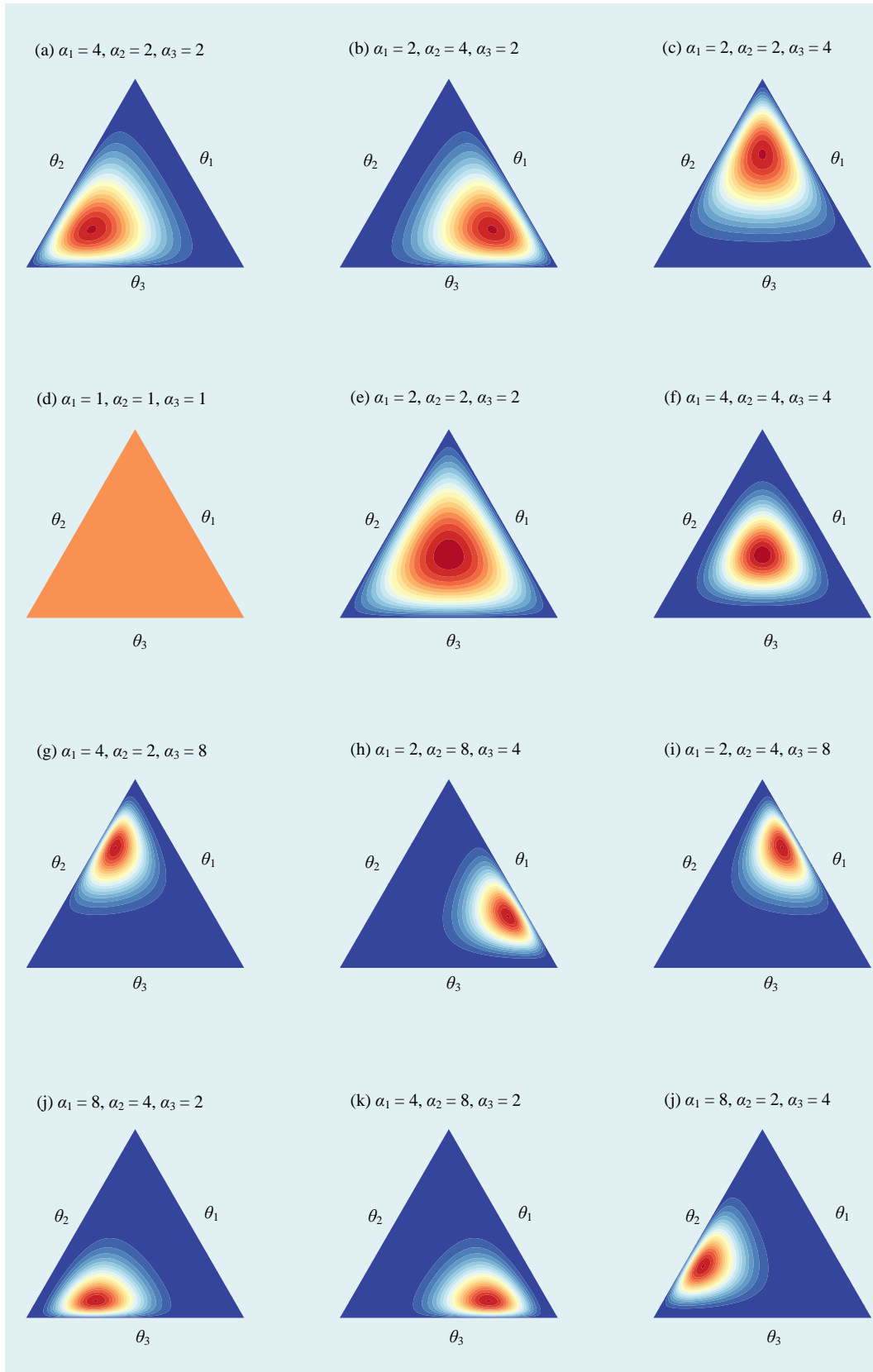


图 20. Dirichlet 分布, 第 2 组 | Bk_2_Ch32_08.ipynb

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

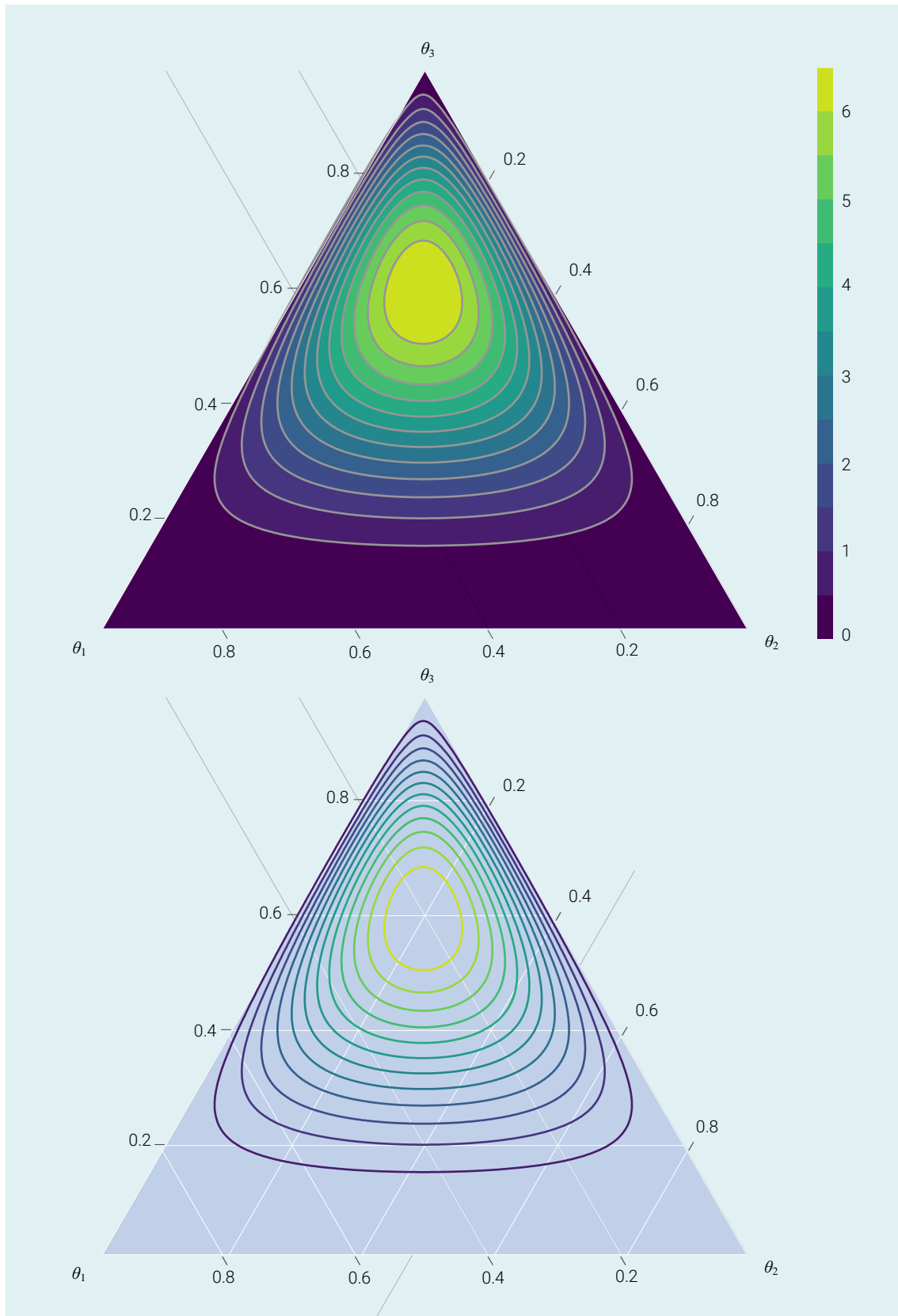



图 21. 用 Plotly 可视化 Dirichlet 分布 |  Bk_2_Ch32_09.ipynb

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com