18 生成扩散模型漫谈（八）：最优扩散方差估计（下）

Aug By 苏剑林 | 2022-08-18 | 42267位读者 引用

在上一篇文章《生成扩散模型漫谈（七）：最优扩散方差估计（上）》中，我们介绍并推导了Analytic-DPM中的扩散模型最优方差估计结果，它是直接给出了已经训练好的生成扩散模型的最优方差的一个解析估计，实验显示该估计结果确实能有效提高扩散模型的生成质量。

这篇文章我们继续介绍Analytic-DPM的升级版，出自同一作者团队的论文《Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models》，在官方Github中被称为“Extended-Analytic-DPM”，下面我们也用这个称呼。

结果回顾

上一篇文章是在DDIM的基础上，推出DDIM的生成过程最优方差应该是

$$\sigma_t^2 + \gamma_t^2 \bar{\sigma}_t^2 \tag{1}$$

其中$\bar{\sigma}_t^2$是分布$p(\boldsymbol{x}_0|\boldsymbol{x}_t)$的方差，它有如下的估计结果（这里取“方差估计2”的结果）：

$$\bar{\sigma}_t^2 = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left(1 - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|^2\right]\right) \tag{2}$$

事后来看，其实估计思路也不算难，假设

$$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) = \frac{1}{\bar{\alpha}_t}\left(\boldsymbol{x}_t - \bar{\beta}_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right) \tag{3}$$

已经准确预测了分布$p(\boldsymbol{x}_0|\boldsymbol{x}_t)$的均值向量，那么根据定义可以得到协方差为

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{x}_t) =& \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t))(\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t))^\top\right] \\ =& \frac{1}{\bar{\alpha}_t^2}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0)(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0)^\top\right] - \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)^\top\end{aligned}$$

两端对$\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)$求平均，以消除对$\boldsymbol{x}_t$的依赖

$$\boldsymbol{\Sigma}_t = \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}[\boldsymbol{\Sigma}(\boldsymbol{x}_t)] = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left(\boldsymbol{I} - \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)^\top\right]\right) \tag{5}$$

最后，对角线元素取平均，使其变为一个标量（或者说协方差是单位阵的倍数），即$\bar{\sigma}_t^2 = \mathrm{Tr}(\boldsymbol{\Sigma}_t)/d$，便可得到估计式$(2)$。

如何改进

在正式介绍Extended-Analytic-DPM之前，我们可以先想想，Analytic-DPM还有什么改进空间？

其实稍加思考就可以发现很多，比如Analytic-DPM假设用来逼近$p(\boldsymbol{x}_0|\boldsymbol{x}_t)$的正态分布协方差矩阵设计为为$\bar{\sigma}_t^2\boldsymbol{I}$，即对角线元素相同的对角阵，那么一个直接的改进就是允许对角线元素互不相同了，即$\mathrm{diag}(\bar{\boldsymbol{\sigma}}_t^2)$，这里约定向量的乘法都是基于Hadamard积进行，比如$\boldsymbol{x}^2 = \boldsymbol{x}\otimes\boldsymbol{x}$。对应的结果就是只考虑$\boldsymbol{\Sigma}_t$的对角线部分，所以从式$(5)$出发，可以得到相应的估计是

$$\bar{\boldsymbol{\sigma}}_t^2 = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left(\boldsymbol{1}_d - \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^2(\boldsymbol{x}_t,t)\right]\right) \tag{6}$$

其中$\boldsymbol{1}_d$是d维全1向量。还有一个更进一步的改进是保留$\bar{\boldsymbol{\sigma}}_t^2$对$\boldsymbol{x}_t$的依赖关系，即考虑$\bar{\boldsymbol{\sigma}}_t^2(\boldsymbol{x}_t)$，这就跟$\boldsymbol{\mu}(\boldsymbol{x}_t)$类似，需要用一个以$\boldsymbol{x}_t$为输入的模型来学习它。

那么可不可以考虑完整的$\boldsymbol{\Sigma}_t$呢？理论上可以，实际上基本不可行，因为完整的$\boldsymbol{\Sigma}_t$是一个$d\times d$矩阵，对于图片场景来说，d是图片的总像素个数，即便是对于cifar10来说也

已经有$d = 32^2 \times 3 = 3072$了，更不用说更高分辨率的图片。所以结合实验背景，$d \times d$矩阵在储存和计算上的成本都过大了。

除此之外，可能有一个问题不少读者都没意识到，就是前面的解析解推导都依赖于$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) = \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}[\boldsymbol{x}_0]$，事实上$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$是由模型学习出来的，它未必能够精确等于均值$\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}[\boldsymbol{x}_0]$，这就是Extended-Analytic-DPM的论文标题所提到的Imperfect Mean的含义。如果在Imperfect Mean下改进估计结果，更加有实践意义。

最大似然

假设均值模型$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$已经事先训练好，那么待定分布$\mathcal{N}(\boldsymbol{x}_0; \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t), \bar{\sigma}_t^2\boldsymbol{I})$的参数就只剩下了$\bar{\sigma}_t^2$，对应的负对数似然为

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[-\log \mathcal{N}(\boldsymbol{x}_0; \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t), \bar{\sigma}_t^2\boldsymbol{I})\right]\\ =&\frac{\mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{x}_0\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)\tilde{p}(\boldsymbol{x}_0)}\left[\left\|\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)\right\|^2\right]}{2\bar{\sigma}_t^2} + \frac{d}{2}\log \bar{\sigma}_t^2 + \frac{d}{2}\log 2\pi\end{aligned}\tag{7}$$

可以解得取最小值正好是

$$\bar{\sigma}_t^2 = \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{x}_0\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)\tilde{p}(\boldsymbol{x}_0)}\left[\left\|\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)\right\|^2\right]\tag{8}$$

它的特点是$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$未必是准确的均值结果，因此式$(4)$的第二个等号不成立，只能成立第一个等号。将式$(3)$代入，得到

$$\bar{\sigma}_t^2 = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2 d}\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}_0),\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\varepsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t\boldsymbol{x}_0 + \bar{\beta}_t\boldsymbol{\varepsilon}, t)\right\|^2\right]\tag{9}$$

当然，这里只分析了协方差矩阵为$\bar{\sigma}_t^2\boldsymbol{I}$的简单情形，我们也可以考虑更一般的对角阵协方差，即$\mathcal{N}(\boldsymbol{x}_0; \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t), \mathrm{diag}(\bar{\boldsymbol{\sigma}}_t^2))$，对应的结果是

$$\bar{\boldsymbol{\sigma}}_t^2 = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}_0),\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left(\boldsymbol{\varepsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t\boldsymbol{x}_0 + \bar{\beta}_t\boldsymbol{\varepsilon}, t)\right)^2\right]\tag{10}$$

条件方差

如果想要得到带条件$\boldsymbol{x}_t$的协方差$\mathrm{diag}(\bar{\boldsymbol{\sigma}}_t^2(\boldsymbol{x}_t))$，那么就相当于每个分量独立计算，结果是免除了$\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}$这一步平均：

$$\bar{\boldsymbol{\sigma}}_t^2(\boldsymbol{x}_t) = \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t))^2\right] = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))^2\right] \quad (11)$$

其中$\boldsymbol{\epsilon}_t = \frac{\boldsymbol{x}_t - \bar{\alpha}_t \boldsymbol{x}_0}{\bar{\beta}_t}$。跟上一篇文章一样，利用

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathop{\mathrm{argmin}}_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{x}}\left[\|\boldsymbol{x} - \boldsymbol{\mu}\|^2\right] \quad (12)$$

得到

$$\begin{aligned}
\bar{\boldsymbol{\sigma}}_t^2(\boldsymbol{x}_t) &= \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))^2\right] \\
&= \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\mathop{\mathrm{argmin}}_{\boldsymbol{g}} \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\left\|(\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))^2 - \boldsymbol{g}\right\|^2\right] \\
&= \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\mathop{\mathrm{argmin}}_{\boldsymbol{g}(\boldsymbol{x}_t)} \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\left\|(\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))^2 - \boldsymbol{g}(\boldsymbol{x}_t)\right\|^2\right] \\
&= \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\mathop{\mathrm{argmin}}_{\boldsymbol{g}(\boldsymbol{x}_t)} \mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{x}_0\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)\tilde{p}(\boldsymbol{x}_0)}\left[\left\|(\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))^2 - \boldsymbol{g}(\boldsymbol{x}_t)\right\|^2\right]
\end{aligned} \quad (13)$$

这就是Extended-Analytic-DPM中学习条件方差的“NPR-DPM”方案。另外，原论文还提了个“SN-DPM”方案，它是基于Perfect Mean假设而不是Imperfect Mean的。然而论文的实验结果却是SN-DPM要优于NPR-DPM，也就是说论文号称自己在解决Imperfect Mean问题，结果实验显示Perfect Mean假设的方案更好，这就反过来说明Perfect Mean假设其实很贴合实践情况，换句话说Imperfect Mean问题可以视为不存在了。

两个阶段

可能读者有疑问，一开始不是说《Improved Denoising Diffusion Probabilistic Models》的可学习方差增加了训练难度吗？那Extended-Analytic-DPM为啥又重新去做可训练的方差模型呢？

我们知道，DDPM提供了方差的两种方案$\sigma_t = \frac{\bar{\beta}_{t-1}}{\bar{\beta}_t}\beta_t$和$\sigma_t = \beta_t$，这两种简单方案的效果其实已经相当不错了。这侧面说明，更精细地调整方差对生成结果的影响不大（至少对于完整的T步扩散是这样），主要的还是$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$的学习，方差只是“锦上添花”的作用。如果将方差视为可学习参数或者模型，跟均值模型$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$一同学习，那么随着训练过程变化的方差就会严重干扰均值模型$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$的学习过程，违反了“$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$为主、方差为辅”的原则。

Extended-Analytic-DPM的聪明之处在于，它提出了两阶段的训练方案，即用原始固定方差的测试训练好均值模型$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$，然后固定该模型，并重用该模型的大部分参数来学一个方差模型，这样一来反而“一举三得”：

> 一、降低了参数量和训练成本；
>
> 二、允许重用已经训练好的均值模型；
>
> 三、训练过程更加稳定。

个人思考

到这里，Extended-Analytic-DPM的介绍就基本完成了。有心的读者可能会感觉到，如果说上一篇Analytic-DPM的结果给人“惊艳”之感，那么这一篇Extended-Analytic-DPM就显得中规中矩，没什么太动人心弦的地方。可以说，Extended-Analytic-DPM就是Analytic-DPM的平凡推广，尽管实验结果显示它还是能带来不错的提升，但总体而言给人的感觉就是很平淡了。当然，大体上是因为Analytic-DPM“珠玉在前”，对比之下才显得它暗淡一些，本身也算是一篇比较扎实的工作。

此外，前面我们已经提到，实验结果显示，基于Perfect Mean假设的SN-DPM，效果要比基于Imperfect Mean假设的NPR-DPM要好，同时这一结果也使得原论文的标题有点

“名不副实”了——既然实验显示Perfect Mean假设的方案更好，反过来意味着Imperfect Mean问题可以视为不存在了。原论文并没有对此结果做进一步的分析和评价，笔者想会不会跟方差估计的有偏性有关？大家知道，直接用“除以n”的公式去估计方差是有偏的，而NPR-DPM正是基于它来操作的，相比之下SN-DPM则是直接取估计二阶矩，二阶矩的估计是无偏的。总感觉有点道理，但也不能完全说通，有点迷～

最后，不知道读者会不会跟笔者一样有个疑问：在给定$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$的前提下，为什么不直接用像式$(7)$的负对数似然为损失函数来学习方差，而是要重新设计NPR-DPM或SN-DPM这两种MSE形式的loss？MSE形式的loss有什么特别的好处吗？笔者暂时也没想到答案。

文章小结

本文介绍了论文Analytic-DPM的升级版——“Extended-Analytic-DPM”中的扩散模型最优方差估计结果，它主要针对不完美均值情形进行了推导，并提出了有条件方差的学习方案。

如果您需要引用本文，请参考：

苏剑林. (Aug. 18, 2022). 《生成扩散模型漫谈（八）：最优扩散方差估计（下）》[Blog post]. Retrieved from https://spaces.ac.cn/archives/9246

@online{kexuefm-9246,
title={生成扩散模型漫谈（八）：最优扩散方差估计（下）},
author={苏剑林},
year={2022},
month={Aug},

```
    url={\url{https://spaces.ac.cn/archives/9246}},
}
```