12 Aug

生成扩散模型漫谈（七）：最优扩散方差估计（上）

By 苏剑林 | 2022-08-12 | 74569位读者 引用

对于生成扩散模型来说，一个很关键的问题是生成过程的方差应该怎么选择，因为不同的方差会明显影响生成效果。

在《生成扩散模型漫谈（二）：DDPM = 自回归式VAE》我们提到，DDPM分别假设数据服从两种特殊分布推出了两个可用的结果；《生成扩散模型漫谈（四）：DDIM = 高观点DDPM》中的DDIM则调整了生成过程，将方差变为超参数，甚至允许零方差生成，但方差为0的DDIM的生成效果普遍差于方差非0的DDPM；而《生成扩散模型漫谈（五）：一般框架之SDE篇》显示前、反向SDE的方差应该是一致的，但这原则上在$\Delta t \to 0$时才成立；《Improved Denoising Diffusion Probabilistic Models》则提出将它视为可训练参数来学习，但会增加训练难度。

所以，生成过程的方差究竟该怎么设置呢？今年的两篇论文《Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models》和《Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models》算是给这个问题提供了比较完美的答案。接下来我们一起欣赏一下它们的结果。

不确定性

事实上，这两篇论文出自同一团队，作者也基本相同。第一篇论文（简称Analytic-DPM）下面简称在DDIM的基础上，推导了无条件方差的一个解析解；第二篇论文（简称Extended-Analytic-DPM）则弱化了第一篇论文的假设，并提出了有条件方差的优化方法。本文首先介绍第一篇论文的结果。

在《生成扩散模型漫谈（四）：DDIM = 高观点DDPM》中，我们推导了对于给定的$p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \bar{\alpha}_t\boldsymbol{x}_0, \bar{\beta}_t^2\boldsymbol{I})$，对应的$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$的一般解为

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}\boldsymbol{x}_t + \gamma_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}\right) \tag{1}$$

其中$\gamma_t = \bar{\alpha}_{t-1} - \frac{\bar{\alpha}_t\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}$，$\sigma_t$就是可调的标准差参数。在DDIM中，接下来的处理流程是：用$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$来估计$\boldsymbol{x}_0$，然后认为

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \approx p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0 = \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)) \tag{2}$$

然而，从贝叶斯的角度来看，这个处理是非常不妥的，因为从$\boldsymbol{x}_t$预测$\boldsymbol{x}_0$不可能完全准确，它带有一定的不确定性，因此我们应该用概率分布而非确定性的函数来描述它。事实上，严格地有

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \int p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) p(\boldsymbol{x}_0|\boldsymbol{x}_t) d\boldsymbol{x}_0 \tag{3}$$

精确的$p(\boldsymbol{x}_0|\boldsymbol{x}_t)$通常是没法获得的，但这里只要一个粗糙的近似，因此我们用正态分布$\mathcal{N}(\boldsymbol{x}_0; \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t), \bar{\sigma}_t^2 \boldsymbol{I})$去逼近它（如何逼近我们稍后再讨论）。有了这个近似分布后，我们可以写出

$$\begin{aligned}
\boldsymbol{x}_{t-1} =&\, \frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}\boldsymbol{x}_t + \gamma_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\varepsilon}_1 \\
\approx&\, \frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}\boldsymbol{x}_t + \gamma_t \left(\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) + \bar{\sigma}_t \boldsymbol{\varepsilon}_2\right) + \sigma_t \boldsymbol{\varepsilon}_1 \\
=&\, \left(\frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}\boldsymbol{x}_t + \gamma_t \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)\right) + \underbrace{\left(\sigma_t \boldsymbol{\varepsilon}_1 + \gamma_t \bar{\sigma}_t \boldsymbol{\varepsilon}_2\right)}_{\sim \sqrt{\sigma_t^2 + \gamma_t^2 \bar{\sigma}_t^2}\boldsymbol{\varepsilon}}
\end{aligned} \tag{4}$$

其中$\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$。可以看到，$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$更加接近均值为$\frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}\boldsymbol{x}_t + \gamma_t \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$、协方差为$\left(\sigma_t^2 + \gamma_t^2 \bar{\sigma}_t^2\right)\boldsymbol{I}$的正态分布，其中均值跟以往的结果

是一致的，不同的是方差多出了$\gamma_t^2\bar{\sigma}_t^2$这一项，因此即便$\sigma_t=0$，对应的方差也不为0。多出来的这一项，就是第一篇论文所提的最优方差的修正项。

均值优化

现在我们来讨论如何用$\mathcal{N}(\boldsymbol{x}_0;\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t),\bar{\sigma}_t^2\boldsymbol{I})$去逼近真实的$p(\boldsymbol{x}_0|\boldsymbol{x}_t)$，说白了就是求出$p(\boldsymbol{x}_0|\boldsymbol{x}_t)$的均值和协方差。

对于均值$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$来说，它依赖于$\boldsymbol{x}_t$，所以需要一个模型来拟合它，而训练模型就需要损失函数。利用

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathop{\text{argmin}}_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{x}}\left[\Vert\boldsymbol{x}-\boldsymbol{\mu}\Vert^2\right] \tag{5}$$

我们得到

$$\begin{aligned}
\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) =&\, \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}[\boldsymbol{x}_0] \\
=&\, \mathop{\text{argmin}}_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\Vert\boldsymbol{x}_0-\boldsymbol{\mu}\Vert^2\right] \\
=&\, \mathop{\text{argmin}}_{\boldsymbol{\mu}(\boldsymbol{x}_t)} \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\Vert\boldsymbol{x}_0-\boldsymbol{\mu}(\boldsymbol{x}_t)\Vert^2\right] \\
=&\, \mathop{\text{argmin}}_{\boldsymbol{\mu}(\boldsymbol{x}_t)} \mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\Vert\boldsymbol{x}_0-\boldsymbol{\mu}(\boldsymbol{x}_t)\Vert^2\right]
\end{aligned} \tag{6}$$

这就是训练$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$所用的损失函数。如果像之前一样引入参数化

$$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) = \frac{1}{\bar{\alpha}_t}\left(\boldsymbol{x}_t - \bar{\beta}_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right) \tag{7}$$

就可以得到DDPM训练所用的损失函数形式$\left\Vert\boldsymbol{\varepsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t \boldsymbol{x}_0 + \bar{\beta}_t \boldsymbol{\varepsilon}, t)\right\Vert^2$了。关于均值优化的结果是跟以往一致的，没有什么改动。

方差估计1

类似地，根据定义，协方差矩阵应该是

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

其中$\boldsymbol{\mu}_0$可以是任意常向量，这对应于协方差的平移不变性。

上式估计的是完整的协方差矩阵，但并不是我们想要的，因为目前我们是想要用$\mathcal{N}(\boldsymbol{x}_0;\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t),\bar{\sigma}_t^2\boldsymbol{I})$去逼近$p(\boldsymbol{x}_0|\boldsymbol{x}_t)$，其中设计的协方差矩阵为$\bar{\sigma}_t^2\boldsymbol{I}$，它有两个特点：

> 1、跟$\boldsymbol{x}_t$无关：为了消除对$\boldsymbol{x}_t$的依赖，我们对全体$\boldsymbol{x}_t$求平均，即$\boldsymbol{\Sigma}_t = \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}[\boldsymbol{\Sigma}(\boldsymbol{x}_t)]$；
>
> 2、单位阵的倍数：这意味着我们只用考虑对角线部分，并且对对角线元素取平均，即$\bar{\sigma}_t^2 = \text{Tr}(\boldsymbol{\Sigma}_t)/d$，其中$d=\dim(\boldsymbol{x})$。

于是我们有

$$\begin{aligned}\bar{\sigma}_t^2 =&\, \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\frac{\Vert\boldsymbol{x}_0 - \boldsymbol{\mu}_0\Vert^2}{d}\right] - \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\frac{\Vert\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0\Vert^2}{d}\right] \\ =&\, \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\Vert\boldsymbol{x}_0 - \boldsymbol{\mu}_0\Vert^2\right] - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\Vert\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0\Vert^2\right] \\ =&\, \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\left[\Vert\boldsymbol{x}_0 - \boldsymbol{\mu}_0\Vert^2\right] - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\Vert\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0\Vert^2\right]\end{aligned}\tag{9}$$

这是笔者给出的关于$\bar{\sigma}_t^2$的一个解析形式，在$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$完成训练的情况下，可以通过采样一批$\boldsymbol{x}_0$和$\boldsymbol{x}_t$来近似计算上式。

特别地，如果取$\boldsymbol{\mu}_0 = \mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}[\boldsymbol{x}_0]$，那么刚好可以写成

$$\bar{\sigma}_t^2 = \mathbb{V}ar[\boldsymbol{x}_0] - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\Vert\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0\Vert^2\right]\tag{10}$$

这里的$\mathbb{V}ar[\boldsymbol{x}_0]$是全体训练数据$\boldsymbol{x}_0$的像素级方差。如果$\boldsymbol{x}_0$的每个像素值都在$[a,b]$区间

内，那么它的方差显然不会超过$\left(\frac{b-a}{2}\right)^2$，从而有不等式

$$\bar{\sigma}_t^2 \leq \mathbb{V}ar[\boldsymbol{x}_0] \leq \left(\frac{b-a}{2}\right)^2 \tag{11}$$

方差估计2

刚才的解是笔者给出的、认为比较直观的一个解，Analytic-DPM原论文则给出了一个略有不同的解，但笔者认为相对来说没那么直观。通过代入式(7)，我们可以得到：

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

此时如果两端对$\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)$求平均，我们有

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\left(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\right)\left(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\right)^{\top}\right] \\
=&\,\mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\right)\left(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\right)^{\top}\right]
\end{aligned} \tag{13}$$

别忘了$p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t;\bar{\alpha}_t\boldsymbol{x}_0, \bar{\beta}_t^2\boldsymbol{I})$，所以$\bar{\alpha}_t\boldsymbol{x}_0$实际上就是$p(\boldsymbol{x}_t|\boldsymbol{x}_0)$的均值，那么$\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\right)\left(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\right)^{\top}\right]$实际上是在求$p(\boldsymbol{x}_t|\boldsymbol{x}_0)$的均值的协方差矩阵，结果显然就是$\bar{\beta}_t^2\boldsymbol{I}$，所以

$$\mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\right)\left(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\right)^{\top}\right] = \mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\left[\bar{\beta}_t^2\boldsymbol{I}\right] = \bar{\beta}_t^2\boldsymbol{I} \quad (1$$

那么

$$\boldsymbol{\Sigma}_t = \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}[\boldsymbol{\Sigma}(\boldsymbol{x}_t)] = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left(\boldsymbol{I} - \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)^{\top}\right]\right) \tag{15}$$

两边取迹然后除以d，得到

$$\bar{\sigma}_t^2 = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left(1 - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\Vert\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\Vert^2\right]\right) \leq \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2} \tag{16}$$

这就得到了另一个估计和上界，这就是Analytic-DPM的原始结果。

实验结果

原论文的实验结果显示，Analytic-DPM所做的方差修正，主要在生成扩散步数较少时会有比较明显的提升，所以它对扩散模型的加速比较有意义：

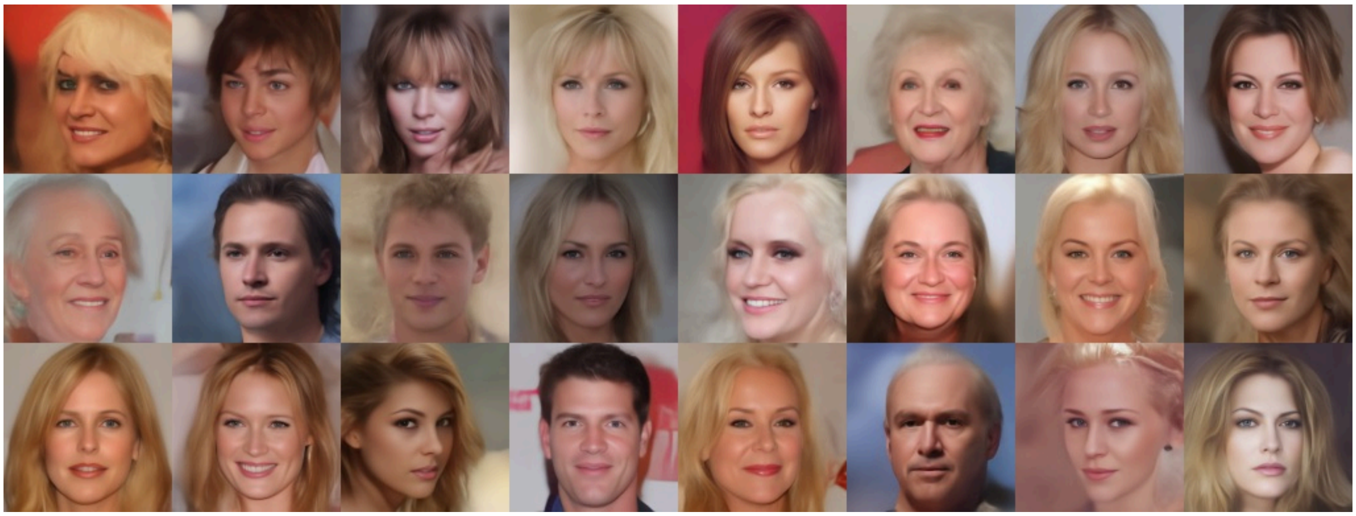
Model \ # timesteps K	10	25	50	100	200	400	1000
CIFAR10 (LS)							
DDPM, $\sigma_n^2 = \tilde{\beta}_n$	44.45	21.83	15.21	10.94	8.23	6.43	5.11
DDPM, $\sigma_n^2 = \beta_n$	233.41	125.05	66.28	31.36	12.96	4.86	†**3.04**
Analytic-DDPM	**34.26**	**11.60**	**7.25**	**5.40**	**4.01**	**3.62**	4.03
DDIM	21.31	10.70	7.74	6.08	5.07	4.61	4.13
Analytic-DDIM	**14.00**	**5.81**	**4.04**	**3.55**	**3.39**	**3.50**	**3.74**
CIFAR10 (CS)							
DDPM, $\sigma_n^2 = \tilde{\beta}_n$	34.76	16.18	11.11	8.38	6.66	5.65	4.92
DDPM, $\sigma_n^2 = \beta_n$	205.31	84.71	37.35	14.81	5.74	**3.40**	**3.34**
Analytic-DDPM	**22.94**	**8.50**	**5.50**	**4.45**	**4.04**	3.96	4.31
DDIM	34.34	16.68	10.48	7.94	6.69	5.78	4.89
Analytic-DDIM	**26.43**	**9.96**	**6.02**	**4.88**	**4.92**	**5.00**	**4.66**
CelebA 64x64							
DDPM, $\sigma_n^2 = \tilde{\beta}_n$	36.69	24.46	18.96	14.31	10.48	8.09	5.95
DDPM, $\sigma_n^2 = \beta_n$	294.79	115.69	53.39	25.65	9.72	**3.95**	**3.16**
Analytic-DDPM	**28.99**	**16.01**	**11.23**	**8.08**	**6.51**	5.87	5.21
DDIM	20.54	13.45	9.33	6.60	4.96	4.15	3.40
Analytic-DDIM	**15.62**	**9.22**	**6.13**	**4.29**	**3.46**	**3.38**	**3.13**

Analytic-DPM主要在扩散步数较少时会有比较明显的效果提升

笔者也在之前自己实现的代码上尝试了Analytic-DPM的修正，参考代码为：

> **Github：https://github.com/bojone/Keras-DDPM/blob/main/adpm.py**

当扩散步数为10时，DDPM与Analytic-DDPM的效果对比如下图：



DDPM在扩散步数为10时的生成结果

Analytic-DDPM在扩散步数为10时的生成结果

可以看到，在扩散步数较小时，DDPM的生成效果比较光滑，有点“重度磨皮”的感觉，相比之下Analytic-DDPM的结果显得更真实一些，但是也带来了额外的噪点。从评价指标来说，Analytic-DDPM要更好一些。

吹毛求疵

至此，我们已经完成了Analytic-DPM的介绍，推导过程略带有一些技巧性，但不算太复杂，至少思路上还是很明朗的。如果读者觉得还是很难懂，那不妨再去看看原论文在附录中用7页纸、13个引理完成的推导，想必看到之后就觉得本文的推导是多么友好了哈哈～

诚然，从首先得到这个方差的解析解来说，我为原作者们的洞察力而折服，但不得不说的是，从“事后诸葛亮”的角度来说，Analytic-DPM在推导和结果上都走了一些的“弯路”，显得“太绕”、”太巧“，从而感觉不到什么启发性。其中，一个最明显的特点是，原论文的结果都用了$\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)$来表达，这就带来了三个问题：一来使得推导过程特别不直观，难以理解“怎么想到的”；二来要求读者额外了解得分匹配的相关结果，增加了理解难度；最后落到实践时，$\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)$又要用回$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$或$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$来表示，多绕一道。

本文推导的出发点是，我们是在估计正态分布的参数，对于正态分布来说，矩估计与最大似然估计相同，因此直接去估算相应的均值方差即可。结果上，没必要强行在形式上去跟$\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)$、得分匹配对齐，因为很明显Analytic-DPM的baseline模型是DDIM，DDIM本身就没有以得分匹配为出发点，增加与得分匹配的联系，于理论和实验都无益。直接跟$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$或$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$对齐，形式上更加直观，而且更容易跟实验形式进行转换。

文章小结

本文分享了论文Analytic-DPM中的扩散模型最优方差估计结果，它给出了直接可用的最优方差估计的解析式，使得我们不需要重新训练就可以直接应用它来改进生成效果。笔者用自己的思路简化了原论文的推导，并进行了简单的实验验证。

如果您需要引用本文，请参考：

苏剑林. (Aug. 12, 2022). 《生成扩散模型漫谈（七）：最优扩散方差估计（上）》[Blog post]. Retrieved from https://spaces.ac.cn/archives/9245

```
@online{kexuefm-9245,
    title={生成扩散模型漫谈（七）：最优扩散方差估计（上）},
    author={苏剑林},
    year={2022},
    month={Aug},
    url={\url{https://spaces.ac.cn/archives/9245}},
}
```