My solutions to

Deep Learning: Foundations and Concepts

Dario Miro Konopatzki

20 Diffusion Models

20.1

Mean

$$\begin{aligned}
\mathbb{E}[Z_t] &= \mathbb{E}\left[\sqrt{1-\beta_t}Z_{t-1} + \sqrt{\beta_t}\mathcal{E}_t\right] \\
&= \sqrt{1-\beta_t}\mathbb{E}[Z_{t-1}] + \sqrt{\beta_t}\mathbb{E}[\mathcal{E}_t] \qquad \text{linearity of } \mathbb{E} \\
&= \sqrt{1-\beta_t}\mathbb{E}[Z_{t-1}] \qquad \mathcal{E}_t \sim \mathcal{N}(0,1) \text{ so i.p. } \mathbb{E}[\mathcal{E}_t] = 0
\end{aligned}$$

$$\begin{aligned}
\|\mathbb{E}[Z_t]\| &= \left\|\sqrt{1-\beta_t}\mathbb{E}[Z_{t-1}]\right\| \\
&= \left|\sqrt{1-\beta_t}\right| \|\mathbb{E}[Z_{t-1}]\| \\
&< \|\mathbb{E}[Z_{t-1}]\| \qquad \left|\sqrt{1-\beta_t}\right| < 1 \text{ since } 0 < \beta_t < 1
\end{aligned}$$

Auxiliary Calculations

$$\begin{aligned}
\mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] &= \mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] - \mathbb{V}[\mathcal{E}_t] + \mathbb{V}[\mathcal{E}_t] \\
&= \mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] - \left(\mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] - \mathbb{E}[\mathcal{E}_t]\mathbb{E}[\mathcal{E}_t]^\top\right) + \mathbb{V}[\mathcal{E}_t] \\
&= \mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] - \mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] + \mathbb{E}[\mathcal{E}_t]\mathbb{E}[\mathcal{E}_t]^\top + \mathbb{V}[\mathcal{E}_t] \\
&= \mathbb{E}[\mathcal{E}_t]\mathbb{E}[\mathcal{E}_t]^\top + \mathbb{V}[\mathcal{E}_t] \\
&= \mathbb{I} \qquad \mathbb{E}[\mathcal{E}_t] = 0,\ \mathbb{V}[\mathcal{E}_t] = \mathbb{I} \text{ since by assumption } \mathcal{E}_t \sim \mathcal{N}(0,\mathbb{I})
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[Z_tZ_t^\top] &= \mathbb{E}\left[\left(\sqrt{1-\beta_t}Z_{t-1} + \sqrt{\beta_t}\mathcal{E}_t\right)\left(\sqrt{1-\beta_t}Z_{t-1} + \sqrt{\beta_t}\mathcal{E}_t\right)^\top\right] \\
&= \mathbb{E}\Big[(1-\beta_t)Z_{t-1}Z_{t-1}^\top + \sqrt{1-\beta_t}\sqrt{\beta_t}Z_{t-1}\mathcal{E}_t^\top
\end{aligned}$$

$$
\begin{aligned}
&\quad + \sqrt{\beta_t}\sqrt{1-\beta_t}\mathcal{E}_t Z_{t-1}^\top + \beta_t \mathcal{E}_t \mathcal{E}_t^\top \Big] \\
&= (1-\beta_t)\mathbb{E}\left[Z_{t-1}Z_{t-1}^\top\right] + \sqrt{1-\beta_t}\sqrt{\beta_t}\mathbb{E}\left[Z_{t-1}\mathcal{E}_t^\top\right] \\
&\quad + \sqrt{\beta_t}\sqrt{1-\beta_t}\mathbb{E}\left[\mathcal{E}_t Z_{t-1}^\top\right] + \beta_t\mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] \qquad \mathbb{E} \text{ linear} \\
&= (1-\beta_t)\mathbb{E}\left[Z_{t-1}Z_{t-1}^\top\right] + \sqrt{1-\beta_t}\sqrt{\beta_t}\mathbb{E}\left[Z_{t-1}\right]\mathbb{E}\left[\mathcal{E}_t^\top\right] \\
&\quad + \sqrt{\beta_t}\sqrt{1-\beta_t}\mathbb{E}\left[\mathcal{E}_t\right]\mathbb{E}\left[Z_{t-1}^\top\right] + \beta_t\mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] \qquad Z_{t-1} \perp \mathcal{E}_t \\
&\overset{(\star)}{=} (1-\beta_t)\mathbb{E}\left[Z_{t-1}Z_{t-1}^\top\right] + \beta_t\mathbb{E}\left[\mathcal{E}_t\mathcal{E}_t^\top\right] \\
&= (1-\beta_t)\mathbb{E}\left[Z_{t-1}Z_{t-1}^\top\right] + \beta_t\mathbb{I}
\end{aligned}
$$

$(\star)$ $\mathcal{E}_t \sim \mathcal{N}(0, \mathbb{I})$, i.p. $\mathbb{E}[\mathcal{E}_t] = 0$

Covariance

$$
\begin{aligned}
\|\mathrm{cov}(Z_t) - \mathbb{I}\| &= \|\mathbb{E}\left[Z_t Z_t^\top\right] - \mathbb{E}[Z_t]\mathbb{E}[Z_t]^\top - \mathbb{I}\| \\
&= \|(1-\beta_t)\mathbb{E}\left[Z_{t-1}Z_{t-1}^\top\right] + \beta_t\mathbb{I} - (1-\beta_t)\mathbb{E}[Z_{t-1}]\mathbb{E}[Z_{t-1}]^\top - \mathbb{I}\| \\
&= \|(1-\beta_t)\left(\mathbb{E}\left[Z_{t-1}Z_{t-1}^\top\right] - \mathbb{E}[Z_{t-1}]\mathbb{E}[Z_{t-1}]^\top - \mathbb{I}\right)\| \\
&= \|(1-\beta_t)\left(\mathrm{cov}(Z_{t-1}) - \mathbb{I}\right)\| \\
&= |1-\beta_t|\|\mathrm{cov}(Z_{t-1}) - \mathbb{I}\| \\
&< \|\mathrm{cov}(Z_{t-1}) - \mathbb{I}\| \qquad |1-\beta_t| < 1 \text{ since } 0 < \beta_t < 1
\end{aligned}
$$

20.2

For every x s.t. $q_X(x) \neq 0$:

$$
\begin{aligned}
q_{Z_1|X=x}(z_1) &= \frac{q_{Z_1,X}(z_1, x)}{q_X(x)} \qquad \text{def. of conditional density} \\
&\overset{(\star)}{=} \frac{q_{\mathcal{E}_1,X}\left(\frac{1}{\sqrt{\beta_1}}\left(z_1 - \sqrt{1-\beta_1}x\right), x\right)}{q_X(x)} \cdot \left|\det\begin{pmatrix} \frac{1}{\sqrt{\beta_1}}\mathbb{I}_D & -\frac{\sqrt{1-\beta_1}}{\sqrt{\beta_1}}\mathbb{I}_D \\ 0 & \mathbb{I}_D \end{pmatrix}\right| \\
&= \frac{1}{\sqrt{\beta_1^D}} \frac{q_{\mathcal{E}_1,X}\left(\frac{1}{\sqrt{\beta_1}}\left(z_1 - \sqrt{1-\beta_1}x\right), x\right)}{q_X(x)} \\
&= \frac{1}{\sqrt{\beta_1^D}} \frac{q_{\mathcal{E}_1}\left(\frac{1}{\sqrt{\beta_1}}\left(z_1 - \sqrt{1-\beta_1}x\right)\right)\cancel{q_X(x)}}{\cancel{q_X(x)}} \qquad \mathcal{E}_1 \perp X \\
&\overset{\dagger}{=} \frac{1}{\sqrt{\beta_1^D}} \frac{1}{\sqrt{(2\pi)^D \mathbb{I}_D}}
\end{aligned}
$$

$$
\begin{aligned}
&\cdot e^{-\frac{1}{2}\left(\frac{1}{\sqrt{\beta_1}}\left(z_1-\sqrt{1-\beta_1}x\right)-0\right)^\top \mathbb{I}_D^{-1}\left(\frac{1}{\sqrt{\beta_1}}\left(z_1-\sqrt{1-\beta_1}x\right)-0\right)} \\
&= \frac{1}{\sqrt{(2\pi)^D\beta_1^D\mathbb{I}_D}} e^{-\frac{1}{2}\left(z_1-\sqrt{1-\beta_1}x\right)^\top \frac{1}{\beta_1}\mathbb{I}_D\left(z_1-\sqrt{1-\beta_1}x\right)} \\
&= \frac{1}{\sqrt{(2\pi)^D\det(\beta_1\mathbb{I}_D)}} e^{-\frac{1}{2}\left(z_1-\sqrt{1-\beta_1}x\right)^\top (\beta_1\mathbb{I}_D)^{-1}\left(z_1-\sqrt{1-\beta_1}x\right)}
\end{aligned}
$$

which is density of distribution $\mathcal{N}\left(\sqrt{1-\beta_1}x, \beta_1\mathbb{I}\right)$.

($\star$) Change of variable with $g(u,v) := (\sqrt{\beta_1}u + \sqrt{1-\beta_1}v, v)$

($\dagger$) $\mathcal{E}_1 \sim \mathcal{N}(0,\mathbb{I})$

20.3

($\star$) Note that if $X \sim \mathcal{N}(\mu,\Sigma)$, $\mathrm{im}(X) \subseteq \mathbb{R}^D$, $a \in \mathbb{R}\setminus\{0\}, b \in \mathbb{R}^D$, then

$$
\begin{aligned}
p_{aX+b}(y) &= p_X\left(\frac{1}{a}(y-b)\right)\left|\det\left(\frac{1}{a}\mathbb{I}_D\right)\right| \\
&= \frac{1}{\sqrt{(2\pi)^D\det(\Sigma)}} e^{-\frac{1}{2}\left(\frac{1}{a}(y-b)-\mu\right)^\top\Sigma^{-1}\left(\frac{1}{a}(y-b)-\mu\right)} \frac{1}{|a^D|} \\
&= \frac{1}{\sqrt{(2\pi)^D\det(\Sigma)(a^D)^2}} e^{-\frac{1}{2}(y-(a\mu+b))^\top\left(\frac{1}{a^2}\Sigma^{-1}\right)(y-(a\mu+b))} \\
&= \frac{1}{\sqrt{(2\pi)^D\det(\Sigma)\det(a^2\mathbb{I}_D)}} e^{-\frac{1}{2}(y-(a\mu+b))^\top\left(a^2\Sigma\right)^{-1}(y-(a\mu+b))} \\
&= \frac{1}{\sqrt{(2\pi)^D\det(a^2\Sigma)}} e^{-\frac{1}{2}(y-(a\mu+b))^\top\left(a^2\Sigma\right)^{-1}(y-(a\mu+b))}
\end{aligned}
$$

is density for $\mathcal{N}(a\mu+b, a^2\Sigma)$.

Induction: $Z_t = \sqrt{\alpha_t}X + \tilde{\mathcal{E}}_t$ with $\tilde{\mathcal{E}}_t \sim \mathcal{N}(0,(1-\alpha_t)\mathbb{I}_D)$ for all t.

$t = 1$

$$
\begin{aligned}
Z_1 &= \sqrt{1-\beta_1}X + \sqrt{\beta_1}\mathcal{E}_1 && \text{by def.} \\
&= \sqrt{\alpha_1}X + \underbrace{\sqrt{1-\alpha_1}\mathcal{E}_1}_{=:\tilde{\mathcal{E}}_1} && \text{def. of } \alpha_1
\end{aligned}
$$

where $\tilde{\mathcal{E}}_1 \sim \mathcal{N}(0,(1-\alpha_1)\mathbb{I}_D)$ holds via ($\star$) since $\mathcal{E}_1 \sim \mathcal{N}(0,\mathbb{I}_D)$ by assumption.

$t \to t+1$

$$\begin{aligned}
Z_{t+1} &= \sqrt{1-\beta_{t+1}} Z_t + \sqrt{\beta_{t+1}} \mathcal{E}_{t+1} && \text{by def.} \\
&= \sqrt{1-\beta_{t+1}} \left(\sqrt{\alpha_t} X + \tilde{\mathcal{E}}_t \right) + \sqrt{\beta_{t+1}} \mathcal{E}_{t+1} && \text{ind. hypothesis} \\
&= \sqrt{(1-\beta_{t+1})\, \alpha_t} X + \sqrt{1-\beta_{t+1}} \tilde{\mathcal{E}}_t + \sqrt{\beta_{t+1}} \mathcal{E}_{t+1} \\
&= \sqrt{\alpha_{t+1}} X + \underbrace{\sqrt{1-\beta_{t+1}} \tilde{\mathcal{E}}_t + \sqrt{\beta_{t+1}} \mathcal{E}_{t+1}}_{:=\tilde{\mathcal{E}}_{t+1}} && \text{def. of } \alpha_{t+1}
\end{aligned}$$

where $\mathcal{E}_{t+1} \sim \mathcal{N}(0, \mathbb{I}_D)$ by assumption and $\tilde{\mathcal{E}}_t \sim \mathcal{N}(0, (1-\alpha_t)\mathbb{I}_D)$ by induction hypothesis. Hence via ($\star$) with (3.212) it holds that

$$\begin{aligned}
\tilde{\mathcal{E}}_{t+1} &\sim \mathcal{N}\left(0, (1-\beta_{t+1})(1-\alpha_t)\mathbb{I}_D + \beta_{t+1}\mathbb{I}_D\right) \\
&= \mathcal{N}(0, (1-\beta_{t+1}-\alpha_{t+1})\mathbb{I}_D + \beta_{t+1}\mathbb{I}_D) \\
&= \mathcal{N}(0, (1-\alpha_{t+1})\mathbb{I}_D)
\end{aligned}$$

Conditional density

$$\begin{aligned}
q_{Z_t|X=x}(z_t) &= \frac{q_{\sqrt{\alpha_t}X+\tilde{\mathcal{E}}_t,X}(z_t, x)}{q_X(x)} \\
&\overset{(\dagger)}{=} \frac{q_{\tilde{\mathcal{E}}_t,X}(z_t - \sqrt{\alpha_t}x, x)}{p_X(x)} \cdot \underbrace{\left| \det \begin{pmatrix} \mathbb{I}_D & -\sqrt{\alpha_t} \\ 0 & \mathbb{I}_D \end{pmatrix} \right|}_{=1} \\
&\overset{(\ddagger)}{=} \frac{q_{\tilde{\mathcal{E}}_t}(z_t - \sqrt{\alpha_t}x)\cancel{q_X(x)}}{\cancel{q_X(x)}} \\
&= \frac{1}{\sqrt{(2\pi)^D \det((1-\alpha_t)\mathbb{I}_D)}} e^{-\frac{1}{2}(z_t-\sqrt{\alpha_t}x)^\top((1-\alpha_t)\mathbb{I}_D)^{-1}(z_t-\sqrt{\alpha_t}x)}
\end{aligned}$$

is density for $\mathcal{N}(\sqrt{\alpha_t}x, (1-\alpha_t)\mathbb{I}_D)$.

($\dagger$) Change of variable with $g(u,v) := \left(\sqrt{\alpha_t}v + u, v\right)$

($\ddagger$) $(\mathcal{E}_t)_t$ is assumed to be 'independent noise', so X independent of $(\mathcal{E}_\tau)_{1 \le t}$ for all t. It follows that X also independent of measurable function $\tilde{\mathcal{E}}_t$ of $(\mathcal{E}_\tau)_{\tau \le t}$.

20.4

By assumption, $0 < \beta_t < 1$ and thus $0 < (1 - \beta_t) < 1$ for all t. Consequently $\alpha_t = \prod_{\tau=1}^{t}(1 - \beta_\tau) \xrightarrow{t\to\infty} 0$ and it follows that

$$\begin{aligned}
q_{Z_t|X=x}(z) &= \frac{1}{\sqrt{(2\pi)^D \det((1-\alpha_t)\mathbb{I}_D)}} e^{-\frac{1}{2}(z_t - \sqrt{\alpha_t}x)^\top((1-\alpha_t)\mathbb{I}_D)^{-1}(z_t-\sqrt{\alpha_t}x)} \\
&= \frac{1}{\sqrt{(2\pi)^D \underbrace{(1-\alpha_t)}_{\longrightarrow 1}{}^D}} e^{-\frac{1}{2}\underbrace{\frac{1}{1-\alpha_t}}_{\longrightarrow 1}(z_t - \underbrace{\sqrt{\alpha_t}x}_{\longrightarrow 0})^\top(z_t - \underbrace{\sqrt{\alpha_t}x}_{\longrightarrow 0})} \\
&\xrightarrow{t\to\infty} \frac{1}{\sqrt{(2\pi)^D}} e^{-\frac{1}{2}z_t^\top z_t}
\end{aligned}$$

which is density for $\mathcal{N}(0, \mathbb{I}_D)$.

20.5

$$\begin{aligned}
\operatorname{cov}[A+B] &= \mathbb{E}\left[(A+B)(A+B)^\top\right] - \mathbb{E}[A+B]\mathbb{E}[A+B]^\top \\
&= \mathbb{E}\left[AA^\top + AB^\top + BA^\top + BB^\top\right] \\
&\quad - \mathbb{E}[A+B]\mathbb{E}[A+B]^\top \\
&= \mathbb{E}\left[AA^\top\right] + \mathbb{E}\left[AB^\top\right] + \mathbb{E}\left[BA^\top\right] + \left[BB^\top\right] \\
&\quad - \left(\mathbb{E}[A] + \mathbb{E}[B]\right)\left(\mathbb{E}[A] + \mathbb{E}[B]\right)^\top \\
&= \mathbb{E}\left[AA^\top\right] + \mathbb{E}\left[AB^\top\right] + \mathbb{E}\left[BA^\top\right] + \left[BB^\top\right] \\
&\quad - \mathbb{E}[A]\mathbb{E}\left[A\right]^\top - \mathbb{E}[A]\mathbb{E}\left[B\right]^\top - \mathbb{E}[B]\mathbb{E}\left[A\right]^\top - \mathbb{E}[B]\mathbb{E}[B]^\top \\
&\overset{(\star)}{=} \mathbb{E}\left[AA^\top\right] \cancel{+\mathbb{E}[A]\mathbb{E}\left[B^\top\right]} \cancel{+\mathbb{E}[B]\mathbb{E}\left[A^\top\right]} + \mathbb{E}\left[BB^\top\right] \\
&\quad - \mathbb{E}[A]\mathbb{E}\left[A\right]^\top \cancel{-\mathbb{E}[A]\mathbb{E}\left[B\right]^\top} - \cancel{\mathbb{E}[B]\mathbb{E}\left[A\right]^\top} - \mathbb{E}[B]\mathbb{E}[B]^\top \\
&= \mathbb{E}\left[AA^\top\right] - \mathbb{E}[A]\mathbb{E}\left[A^\top\right] + \mathbb{E}\left[BB^\top\right] - \mathbb{E}[B]\mathbb{E}\left[B^\top\right] \\
&= \operatorname{cov}[A] + \operatorname{cov}[B]
\end{aligned}$$

$(\star)$ $A \perp B$

$$\begin{aligned}
\operatorname{cov}(\lambda A) &= \mathbb{E}\left[\lambda A(\lambda A)^\top\right] - \mathbb{E}\left[\lambda A\right]\mathbb{E}[\lambda A]^\top \\
&= \lambda^2\left(\mathbb{E}\left[AA^\top\right] - \mathbb{E}[A]\mathbb{E}[A]^\top\right) \\
&= \lambda^2 \operatorname{cov}(A)
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}[Z_t] &= \mathbb{E}\left[\sqrt{1-\beta_t} Z_{t-1} + \sqrt{\beta_t}\mathcal{E}_t\right] \\
&= \sqrt{1-\beta_t}\mathbb{E}[Z_{t-1}] + \sqrt{\beta_t}\mathbb{E}[\mathcal{E}_t] \\
&= \sqrt{1-\beta_t}\cdot 0 + \sqrt{\beta_t}\cdot 0 \qquad \mathcal{E}_t \sim \mathcal{N}(0,\mathbb{I}) \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{cov}[Z_t] &= \operatorname{cov}\left[\sqrt{1-\beta_t} Z_{t-1} + \sqrt{\beta_t}\mathcal{E}_t\right] \\
&\overset{(\dagger)}{=} \left(\sqrt{1-\beta_t}\right)^2 \operatorname{cov}\left[Z_{t-1}\right] + \left(\sqrt{\beta_t}\right)^2 \operatorname{cov}\left[\mathcal{E}_t\right] \\
&= (1-\beta_t)\,\mathbb{I} + \beta_t\mathbb{I} \qquad \mathcal{E}_t \sim \mathcal{N}(0,\mathbb{I}) \\
&= \mathbb{I} - \beta_t\mathbb{I} + \beta_t\mathbb{I} \\
&= \mathbb{I}
\end{aligned}
$$

($\dagger$) properties of cov as shown above

20.8

$$
\begin{aligned}
&\mathcal{L}(w) + \mathrm{KL}\left(q_Z(z)\|p_{Z|X=x,W=w}(z)\right) \\
&\quad= \int q_Z(z)\ln\left(\frac{p_{X,Z|W=w}(x,z)}{q_Z(z)}\right)dz - \int q_Z(z)\ln\left(\frac{p_{Z|X=x,W=w}(z)}{q_Z(z)}\right)dz \\
&\quad= \int q_Z(z)\left(\ln\left(\frac{p_{X,Z|W=w}(x,z)}{q_Z(z)}\right) - \ln\left(\frac{p_{Z|X=x,W=w}(z)}{q_Z(z)}\right)\right)dz \\
&\quad= \int q_Z(z)\ln\left(\frac{p_{X,Z|W=w}(x,z)}{\cancel{q_Z(z)}}\frac{\cancel{q_Z(z)}}{p_{Z|X=x,W=w}(z)}\right)dz \\
&\quad= \int q_Z(z)\ln\left(\frac{\cancel{p_{Z|X=x,W=w}(z)}p_{X|W=w}(x)}{\cancel{p_{Z|X=x,W=w}(z)}}\right)dz \\
&\quad= \ln p_{X|W=w}(x)\underbrace{\int q_Z(z)dz}_{=1} \\
&\quad= \ln p_{X|W=w}(x)
\end{aligned}
$$