REFERENCES

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024a.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024b.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024.

Stefano Bruno, Ying Zhang, Dong-Young Lim, Ömer Deniz Akyildiz, and Sotirios Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

Jinyuan Chang, Zhao Ding, Yuling Jiao, Ruoxuan Li, and Jerry Zhijian Yang. Deep conditional generative learning: Model and error analysis. *arXiv preprint arXiv:2402.01460*, 2024.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023d.

Sitan Chen, Giannis Daras, and Alexandros G. Dimakis. Restoration-degradation beyond linear diffusions: a non-asymptotic analysis for ddim-type samplers. In *Proceedings of the 40th International Conference on Machine Learning*, 2023e.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Xiuyuan Cheng, Jianfeng Lu, Yixin Tan, and Yao Xie. Convergence of flow-based generative models via proximal gradient descent in wasserstein space. *arXiv preprint arXiv:2310.17582*, 2023.

Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting. *arXiv preprint arXiv:2409.13074*, 2024.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14367–14376, October 2021.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, 2022a.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster accelerating conditional diffusion models for inverse problems through stochastic contraction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

Frank Cole and Yulong Lu. Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian distributions. In *The Twelfth International Conference on Learning Representations*, 2024.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

F. Croitoru, V. Hondru, R. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(9):10850–10869, Sep 2023.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709, 2021.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9935–9946, June 2023.

John Torjus Flåm. The linear model under gaussian mixture inputs: Selected problems in communications, 2013.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *arXiv preprint arXiv:2311.11003*, 2023.

Yuan Gao, Jian Huang, Yuling Jiao, and Shurong Zheng. Convergence of continuous normalizing flows for learning probability distributions. *arXiv preprint arXiv:2404.00551*, 2024.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Shivam Gupta, Ajil Jalal, Aditya Parulekar, Eric Price, and Zhiyang Xun. Diffusion posterior sampling is computationally intractable. *arXiv preprint arXiv:2402.12727*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *arXiv preprint arXiv:2404.09730*, 2024.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. In *Advances in Neural Information Processing Systems*, 2021.

Yuling Jiao, Yanming Lai, Yang Wang, and Bokai Yan. Convergence analysis of flow matching in latent space with transformers. *arXiv preprint arXiv:2404.02538*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201, pp. 946–985, 2023.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*, 2024a.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024b.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024c.

Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement – a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Non-asymptotic convergence of discrete-time diffusion models: New approach and improved rate. *arXiv preprint arXiv:2402.13901*, 2024.

Junlong Lyu, Zhitang Chen, and Shoubo Feng. Sampling is as easy as keeping the consistency: convergence guarantee for consistency models, 2024.

Song Mei and Yuchen Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv preprint arXiv:2309.11420*, 2023.

Brian B. Moser, Arundhati S. Shanbhag, Federico Raue, Stanislav Frolov, Sebastián M. Palacio, and Andreas Dengel. Diffusion models, image super-resolution and everything: A survey. *arXiv preprint arXiv:2401.00736*, 2024.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1530–1538, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the DDPM objective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2256–2265, 2015.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023a.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *arXiv preprint arXiv:2403.01639*, 2024.

Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *arXiv preprint arXiv:2403.17042*, 2024.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*, 2024.

Appendix

A Full List of Notations

For any two functions $f(d,\delta,T)$ and $g(d,\delta,T)$, we write $f(d,\delta,T) \lesssim g(d,\delta,T)$ (resp. $f(d,\delta,T) \gtrsim g(d,\delta,T)$) for some universal constant (not depending on δ, d or T) $L < \infty$ (resp. $L > 0$) if $\limsup_{T\to\infty} |f(d,\delta,T)/\, g(d,\delta,T)| \le L$ (resp. $\liminf_{T\to\infty} |f(d,\delta,T)/g(d,\delta,T)| \ge L$). We write $f(d,\delta,T) \asymp g(d,\delta,T)$ when both $f(d,\delta,T) \lesssim g(d,\delta,T)$ and $f(d,\delta,T) \gtrsim g(d,\delta,T)$ hold. Note that the dependence on δ and d is retained with $\lesssim, \gtrsim, \asymp$. We write $f(d,\delta,T) = O(g(T))$ (resp. $f(d,\delta,T) = \Omega(g(T))$) if $f(d,\delta,T) \lesssim L(d,\delta)g(T)$ (resp. $f(d,\delta,T) \gtrsim L(d,\delta)g(T)$) holds for some $L(d,\delta)$ (possibly depending on δ and d). We write $f(d,\delta,T) = o(g(T))$ if $\limsup_{T\to\infty} |f(d,\delta,T)/g(T)| = 0$. We write $f(d,\delta,T) = \tilde{O}(g(T))$ if $f(d,\delta,T) = O(g(T)(\log g(T))^k)$ for some constant k. Note that the big-O notation omits the dependence on δ and d. In the asymptotic when $\varepsilon^{-1} \to \infty$, we write $f(d,\varepsilon^{-1}) = \mathcal{O}(g(d,\varepsilon^{-1}))$ if $f(d,\delta,\varepsilon^{-1}) \lesssim g(d,\delta,\varepsilon^{-1})(\log g(\varepsilon^{-1}))^k$ for some constant k. Unless otherwise specified, we write $x^i (1 \le i \le d)$ as the i-th element of a vector $x \in \mathbb{R}^d$ and $[A]^{ij}$ as the (i,j)-th element of a matrix A. For a function $f(x) : \mathbb{R}^d \to \mathbb{R}$, we write $\partial_i f(z)$ as a shorthand for $\frac{\partial}{\partial x^i} f(x)\Big|_{x=z}$, and similarly for higher moments. For a vector (resp. matrix), all norms, if not explicitly specified, are referred to 2-norm (resp. spectral norm). For a vector x and matrix P, define $\|x\|_P := \sqrt{a^\intercal P a}$. For matrices A, B, $\mathrm{Tr}(A)$ is the trace of A, and $A \preceq B$ means that $B - A$ is positive semi-definite. For a positive integer n, $[n] := \{1,\dots,n\}$.

B Related Works on Unconditional DDPM Samplers

Given time-averaged L^2 unconditional score estimation error (Hyvärinen, 2005), polynomial-time convergence guarantees have been established for wide families of target distributions (De Bortoli et al., 2021; Chen et al., 2023d; Lee et al., 2023; Chen et al., 2023a; Benton et al., 2024a; Pedrotti et al., 2023; Conforti et al., 2023). For all target distributions with finite second moment, under L^2 score estimation error, $\mathcal{O}(d\log(1/\delta)^2/\varepsilon^2)$ number of steps are sufficient to achieve ε^2 KL divergence between the δ-perturbed target distribution and the generated distribution using the specially designed exponential-decay-then-constant step-sizes (Benton et al., 2024a; Conforti et al., 2023). The analysis usually involves applying the Girsanov change-of-measure framework and the Fokker-Plank equation (Chen et al., 2023d;a) to either the original SDE diffusion process or some transformed process (Benton et al., 2024a; Conforti et al., 2023), followed by an analysis of the discretization of the continuous-time process. More recently, similar convergence guarantees have been established using non-SDE-type techniques, such as with typical sets (Li et al., 2024c) and with tilting factor representations (Liang et al., 2024). Here the new analysis introduced in Liang et al. (2024) is applicable to a larger set of step-sizes (equivalently, noise schedules) than the ones commonly used in previous analyses (Chen et al., 2023a; Benton et al., 2024a; Conforti et al., 2023). In this paper, we employ the same analytical framework as in Liang et al. (2024).

Some other works analyzed sampling errors using a different measure (the Wasserstein-2 distance) (Bruno et al., 2023; Gao et al., 2023; Gao & Zhu, 2024). Beyond stochastic samplers, another line of studies provided theoretical guarantees for the deterministic sampler corresponding to DDPM (Chen et al., 2023e;c; Huang et al., 2024). Besides, Cheng et al. (2023); Benton et al. (2024b); Jiao et al. (2024); Gao et al. (2024) provided guarantees for the closely-related flow-matching model, which learns a deterministic coupling between any two distributions. Also, Lyu et al. (2024); Li et al. (2024b) provided convergence guarantees for the closely-related consistency models (Song et al., 2023b). Finally, in order to achieve an end-to-end analysis, several works also developed sample complexity bounds to achieve the L^2 score estimation error for a variety of distributions (Oko et al., 2023; Shah et al., 2023; Gatmiry et al., 2024; Chen et al., 2024; Cole & Lu, 2024; Zhang et al., 2024; Mei & Wu, 2023; Chen et al., 2023b).

C Details of Numerical Simulations

In Figure 1, we compared the performances of our optimal BO-DDNM sampler (with the $f^*_{t,y}$ in (10)) against the DDNM and DDNM$^+$ samplers (Wang et al., 2023) at different levels of σ_y^2. For Gaussian, we use $\mu_0 = 0$, $d = 4$, $p = 2$, and $y = (0.5 \quad 0.5)$. We first randomly generate a positive definite matrix Σ_0 and uniformly sample $\rho \in [0.4, 0.7)$, and then this correlation coefficient is enforced for any $[\Sigma_0]^{ij}$ where $i \in [p]$ and $j \in \{p+1, \ldots, d\}$. We use the noise schedule in (8) with $c = 3$ and $\delta = 0.0001$ for Gaussian Q_0. For Gaussian mixture, we use $N = 2$, $d = 2$, $p = 1$, and $y = 1$. We set $\pi_n = (0.4 \quad 0.6)$, $\text{diag}(\Sigma_0) = (0.1 \quad 1)$, and $\rho = 0.6$. We further uniformly sample $\{\mu_{0,n}\}_{n=1}^N$ in the space $[-1, 1) \times [-1, 1)$. We use the noise schedule in (8) with $c = 4$ and $\delta = 0.02$ for Gaussian mixture Q_0. We use 150000 samples to estimate the divergence when Q_0 is Gaussian mixture.

In Figure 2, we numerically verify the exact bias in KL divergence as a function of y and ρ for Gaussian Q_0. Here $Q_0 = \mathcal{N}(0, \Sigma_0)$, $d = 4$ and $p = 2$. Suppose that $\sigma_y^2 = 0$. We assume that each element of y has equal values. The correlation coefficient ρ is enforced for any pair of x^i and x^j where $i \in [p]$ and $j \in \{p+1, \ldots, d\}$. We first randomly generate a positive definite matrix Σ_0 and then enforce the correlation condition for any x^i and x^j where $i \in [p]$ and $j \in \{p+1, \ldots, d\}$. We use a sufficiently large number of steps $T = 20000$. The conditional sampler is set as $f_{t,y} = f^*_{t,y}$ given in (10). The noise schedule in (8) with $c = 3$ and $\delta = 0.0001$ is used.

D Derivation of Score Bias for Existing Zero-shot DDPM Samplers

In this section we show some examples of zero-shot conditional samplers proposed in the literature and in particular how they are related to the formulation of interest in (6). We recall the notations H, y, and σ_t from Section 2.3. Also denote

$$\mu_t := \frac{1}{\sqrt{\alpha_t}} x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}} \nabla \log q_t(x_t) = \mathbb{E}_{X_{t-1} \sim Q_{t-1|t}(\cdot|x_t)}[X_{t-1}|x_t]$$

which is the mean of the unconditional reverse-step at time $t \geq 1$.

D.1 Come-Closer-Diffuse-Faster (CCDF)

We first examine the Come-Closer-Diffuse-Faster (CCDF) algorithm (Chung et al., 2022b). The CCDF algorithm using DDPM samplers gives that

$$\begin{aligned} x'_{t-1} &= \mu_t + \sigma_t z_{t,1}, \\ x_{t-1} &= (I - H^\dagger H) x'_{t-1} + \sqrt{\bar{\alpha}_t} H^\dagger y + \sqrt{1-\bar{\alpha}_t} z_{t,2} \end{aligned}$$

where $z_{t,1}, z_{t,2} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ are standard Gaussian random variables. Thus, the conditional mean of the update is

$$\begin{aligned} \mu_{t,y} &= (I - H^\dagger H)\mu_t + \sqrt{\bar{\alpha}_{t-1}} H^\dagger y \\ &= \frac{1}{\sqrt{\alpha_t}} x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}} (I - H^\dagger H) \nabla \log q_t(x_t) + \sqrt{\bar{\alpha}_t} H^\dagger y - \frac{1}{\sqrt{\alpha_t}} H^\dagger H x_t \end{aligned}$$

in which

$$f_{t,y}(x_t) = \frac{1}{1-\alpha_t} \left(\sqrt{\alpha_t}\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H x_t \right).$$

D.2 DDNM and DDNM$^+$

Next, we examine the DDNM algorithm and its modified version DDNM$^+$ (Wang et al., 2023). We first note that the unconditional DDPM satisfies that (cf. (Ho et al., 2020, Equations (7) and (11))),

$$\begin{aligned} \mu_t &:= \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} x_{0|t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t, \\ x_{0|t} &:= \frac{1}{\sqrt{\bar{\alpha}_t}} x_t + \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}} \nabla \log q_t(x_t) = \mathbb{E}_{X_0 \sim Q_{0|t}(\cdot|x_t)}[X_0|x_t]. \end{aligned} \tag{11}$$

Combining these two lines, we have $\mu_t = \frac{1}{\sqrt{\alpha_t}}(x_t + (1-\alpha_t)\nabla \log q_t(x_t))$. In DDNM, $x_{0|t}$ is projected along the direction of the given y, which yields

$$x_{0|t,y} := H^\dagger y + (I_d - H^\dagger H)x_{0|t},$$

and the corresponding conditional mean of the update becomes

$$\mu_{t,y} = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_{0|t,y} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t.$$

Thus,

$$\begin{aligned}
\mu_{t,y} &= \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\left(H^\dagger y + (I_d - H^\dagger H)x_{0|t}\right) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \\
&\overset{(i)}{=} (I_d - H^\dagger H)\mu_t + H^\dagger\left(\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}y + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}Hx_t\right) \\
&= \frac{1}{\sqrt{\alpha_t}}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}(I_d - H^\dagger H)\nabla \log q_t(x_t) \\
&\quad + \left(\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}H^\dagger y + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}H^\dagger H x_t - \frac{1}{\sqrt{\alpha_t}}H^\dagger H x_t\right)
\end{aligned}$$

where (i) follows from (11). Thus, to express this conditional mean in the form of (6),

$$\begin{aligned}
f_{t,y}(x_t) &= \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}H^\dagger y + \frac{1}{1-\alpha_t}\left(\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} - 1\right)H^\dagger H x_t \\
&= \frac{1}{1-\bar{\alpha}_t}\left(\sqrt{\bar{\alpha}_t}H^\dagger y - H^\dagger H x_t\right).
\end{aligned}$$

Here note that $f_{t,y}(x_t)$ is supported on $\mathrm{range}(H^\dagger H)$. Also note that for DDNM, $f_{t,y} = f^*_{t,y}$, which is the BO-DDNM sampler defined in (10), when there is no measurement noise (i.e., $\sigma_y^2 = 0$).

Next we investigate its modified version, DDNM$^+$, in particular when $H = (I_p \quad 0)$. To relate the notations of (Wang et al., 2023, Section 3.3 and Appendix I) with ours, note that $\Sigma = A = H$, $U = I_p$, $V = I_d$, $s_1, \ldots, s_p = 1$, $s_{p+1}, \ldots, s_d = 0$, and $a = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}$. If $\sigma_t \geq a\sigma_y$, we have

$$\Sigma_t = I_d, \quad \Phi_t = \begin{pmatrix}(\sigma_t^2 - a^2\sigma_y^2)I_p & 0 \\ 0 & \sigma_t^2 I_{d-p}\end{pmatrix}.$$

Otherwise, if $\sigma_t < a\sigma_y$, we have

$$\Sigma_t = \begin{pmatrix}\frac{\sigma_t}{a\sigma_y}I_p & 0 \\ 0 & I_{d-p}\end{pmatrix}, \quad \Phi_t = \begin{pmatrix}0 & 0 \\ 0 & \sigma_t^2 I_{d-p}\end{pmatrix}.$$

Observe that the only difference is on the space that supports $H^\dagger H$.

From (Wang et al., 2023, Equations (17) and (18)), we can write

$$\hat{x}_{0|t,y} := (I_d - H^\dagger H)x_{0|t} + \underbrace{\Sigma_t H^\dagger y + (I_d - \Sigma_t)H^\dagger H x_{0|t}}_{\text{supported on } \mathrm{range}(H^\dagger H)},$$

Thus, with similar arguments above,

$$\begin{aligned}
\mu_{t,y} &= \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\hat{x}_{0|t,y} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \\
&= \frac{1}{\sqrt{\alpha_t}}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}(I_d - H^\dagger H)\nabla \log q_t(x_t) \\
&\quad + \frac{1-\alpha_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\underbrace{\left(\sqrt{\bar{\alpha}_t}(\Sigma_t H^\dagger y + (I_d - \Sigma_t)H^\dagger H x_{0|t}) - H^\dagger H x_t\right)}_{\text{supported on } \mathrm{range}(H^\dagger H)}
\end{aligned}$$

where $f_{t,y}$ is again supported on $\mathrm{range}(H^\dagger H)$.

D.3 SAMPLERS USING HIGHER-ORDER DERIVATIVES

Before we end this section, we note that the formulation in (6) only uses (estimates of) first-order derivatives of (unconditional) log-p.d.f.s (a.k.a. unconditional score functions). This might not correspond to the optimal zero-shot sampler, and in practice there have been methods that use both first- and second-order derivatives (namely, in $\partial x_{0|t}(x_t)/\partial x_t$) to achieve better zero-shot sampling performance (Chung et al., 2023; Song et al., 2023a). Nevertheless, the second-order derivatives might be hard to obtain, which require extra machine time and memory in the calculation. We leave investigations to use second-order derivatives in zero-shot conditional samplers as future work.

E PROOF SKETCH OF THEOREM 1

We now provide a proof sketch of Theorem 1 to describe the idea of our analysis approach. The main technical challenge due to mismatched scores is to analyze the expected tilting factor under a mean-perturbed Gaussian, providing an upper bound of the asymptotic orders of all Gaussian non-centralized moments. See the full proof in Appendix F.

To begin, with Lemma 1, we decompose the total error as $\mathrm{KL}(\tilde{Q}_0\|\widehat{P}_0) \le \mathbb{E}_{X_T\sim\tilde{Q}_T}\left[\log\frac{\tilde{q}_T(X_T)}{\widehat{p}_T(X_T)}\right] + \sum_{t=1}^T \mathbb{E}_{X_t,X_{t-1}\sim\tilde{Q}_{t,t-1}}\left[\log\frac{p_{t-1|t}(X_{t-1}|X_t)}{\widehat{p}_{t-1|t}(X_{t-1}|X_t)}\right] + \sum_{t=1}^T \mathbb{E}_{X_t,X_{t-1}\sim\tilde{Q}_{t,t-1}}\left[\log\frac{p_{t-1|t}(X_{t-1}|X_t)}{\widehat{p}_{t-1|t}(X_{t-1}|X_t)}\right]$. These three terms correspond respectively to the *initialization error*, *estimation error*, and *reverse-step error*. The initialization error can be bounded by $\bar{\alpha}_T d$ in order using (Liang et al., 2024, Lemma 3) under Assumption 1. Below we focus on the remaining two terms.

Step 1: Bounding estimation error under mismatch (Lemma 2). At each time $t = 1, \dots, T$, $\log(p_{t-1|t}(x_{t-1}|x_t)/\widehat{p}_{t-1|t}(x_{t-1}|x_t))$ has an explicit expression since they are conditional Gaussians with the same variance. However, differently from the typical matched case, the mean of $P_{t-1|t}$ (i.e., $\mu_t(x_t)$) is no longer equal to the posterior mean of $\tilde{Q}_{t-1|t}$ (i.e., $m_t(x_t)$). Their difference is contained in $\Delta_t(x_t)$, whose asymptotic order needs to be upper-bounded in light of Assumption 2.

Step 2: Decomposing reverse-step error under mismatch (Equation (15)). First we decompose the tilting factor as $\tilde{\zeta}_{t,t-1}(x_t, x_{t-1}) = \tilde{\zeta}_{\text{mis}}(x_t, x_{t-1}) + \tilde{\zeta}_{\text{van}}(x_t, x_{t-1})$, where

$$\tilde{\zeta}_{\text{mis}} := \sqrt{\alpha_t}\Delta_t(x_t)^\intercal(x_{t-1} - m_t(x_t))$$

$$\tilde{\zeta}_{\text{van}} := (\nabla \log \tilde{q}_{t-1}(m_t(x_t)) - \sqrt{\alpha_t}\nabla \log \tilde{q}_t(x_t))^\intercal(x_{t-1} - m_t(x_t)) + \sum_{p=2}^{\infty} T_p(\log \tilde{q}_{t-1}, x_t, m_t(x_t)).$$

Here $\tilde{\zeta}_{\text{mis}}$ captures the factor that contributes to the total bias within $\tilde{\zeta}_{t,t-1}$. Define the oracle sampling process as $\tilde{P}_{t-1|t} = \mathcal{N}(m_{t,y}, \sigma_t^2 I_d)$. Then, the reverse-step error can be decomposed as

$$\begin{aligned}\mathbb{E}_{\tilde{Q}_{t,t-1}}[\tilde{\zeta}_{t,t-1}] - \mathbb{E}_{\tilde{Q}_t\times P_{t-1|t}}[\tilde{\zeta}_{t,t-1}] &= \underbrace{\left(\mathbb{E}_{\tilde{Q}_{t,t-1}}[\tilde{\zeta}_{t,t-1}] - \mathbb{E}_{\tilde{Q}_t\times P_{t-1|t}}[\tilde{\zeta}_{t,t-1}]\right)}_{\mathcal{W}_{\text{oracle, rev-step}}} \\ &+ \underbrace{\left(\mathbb{E}_{\tilde{Q}_t\times \tilde{P}_{t-1|t}}[\tilde{\zeta}_{\text{mis}}] - \mathbb{E}_{\tilde{Q}_t\times P_{t-1|t}}[\tilde{\zeta}_{\text{mis}}]\right)}_{\mathcal{W}_{\text{bias, rev-step}}} + \underbrace{\left(\mathbb{E}_{\tilde{Q}_t\times \tilde{P}_{t-1|t}}[\tilde{\zeta}_{\text{van}}] - \mathbb{E}_{\tilde{Q}_t\times P_{t-1|t}}[\tilde{\zeta}_{\text{van}}]\right)}_{\mathcal{W}_{\text{vanish, rev-step}}}.\end{aligned}$$

Step 3: Bounding $\mathcal{W}_{\text{oracle, rev-step}}$ and $\mathcal{W}_{\text{bias, rev-step}}$ (Equations (16) and (17)). Under Assumption 3, the dominant term of $\mathcal{W}_{\text{oracle, rev-step}}$ is given by (Liang et al., 2024, Theorem 1). Also, the calculation of $\mathcal{W}_{\text{bias, rev-step}}$ is reduced to the difference in conditional mean, which is proportional to $\|\Delta_t(x_t)\|^2$.

Step 4: Bounding $\mathcal{W}_{\text{vanish, rev-step}}$ (Lemmas 3 and 4). To upper-bound $\mathcal{W}_{\text{vanish, rev-step}}$, with results on the matched case in Liang et al. (2024), we need only to characterize the mean of $\tilde{\zeta}_{\text{van}}$ under the mismatched posterior $P_{t-1|t}$. We determine the dominant order in the expected values of all Taylor polynomials, which includes calculating all non-centralized moments. We first calculate the first three non-centralized moments (Lemma 3) and then determine the asymptotic order of all higher moments (Lemma 4). With these, we can finally locate the terms of dominating order in $\mathcal{W}_{\text{vanish, rev-step}}$.

F Proof of Theorem 1 and Corollary 1

Overall, the structure of the proof of Theorem 1 is similar to that for (Liang et al., 2024, Theorem 1). To start, we note that with similar arguments in (Liang et al., 2024, Equation 13), an upper bound on $\mathrm{KL}(\tilde{Q}_0\|\widehat{P}_0)$ is given by

$$\begin{aligned}
&\mathrm{KL}(\tilde{Q}_0\|\widehat{P}_0) \\
&= \mathrm{KL}(\tilde{Q}_T\|\widehat{P}_T) + \sum_{t=1}^{T}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\mathrm{KL}(\tilde{Q}_{t-1|t}(\cdot|X_t)\|\widehat{P}_{t-1|t}(\cdot|X_t))\right] \\
&\quad - \sum_{t=1}^{T}\mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1}}\left[\mathrm{KL}(\tilde{Q}_{t|t-1}(\cdot|X_{t-1})\|\widehat{P}_{t|t-1}(\cdot|X_{t-1}))\right] \\
&\le \mathrm{KL}(\tilde{Q}_T\|\widehat{P}_T) + \sum_{t=1}^{T}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\mathrm{KL}(\tilde{Q}_{t-1|t}(\cdot|X_t)\|\widehat{P}_{t-1|t}(\cdot|X_t))\right] \\
&= \underbrace{\mathbb{E}_{X_T\sim\tilde{Q}_T}\left[\log\frac{\tilde{q}_T(X_T)}{\widehat{p}_T(X_T)}\right]}_{\text{Term 1: initialization error}} + \underbrace{\sum_{t=1}^{T}\mathbb{E}_{X_t,X_{t-1}\sim\tilde{Q}_{t,t-1}}\left[\log\frac{p_{t-1|t}(X_{t-1}|X_t)}{\widehat{p}_{t-1|t}(X_{t-1}|X_t)}\right]}_{\text{Term 2: estimation error}} \\
&\quad + \underbrace{\sum_{t=1}^{T}\mathbb{E}_{X_t,X_{t-1}\sim\tilde{Q}_{t,t-1}}\left[\log\frac{\tilde{q}_{t-1|t}(X_{t-1}|X_t)}{p_{t-1|t}(X_{t-1}|X_t)}\right]}_{\text{Term 3: reverse-step error}}.
\end{aligned} \tag{12}$$

The last equality holds because $\widehat{p}_T = p_T$. Now, we provide an upper bound for the reverse-step error that is ready for further analysis. In the following lemma, we show that the mismatched $\tilde{q}_{t-1|t}$ is an exponentially tilted form of $p_{t-1|t}$.

Lemma 1. *Fixed $t \ge 1$. For any fixed $x_t \in \mathbb{R}^d$, as long as $\tilde{q}_{t-1}$ exists, we have*

$$\tilde{q}_{t-1|t}(x_{t-1}|x_t) = \frac{p_{t-1|t}(x_{t-1}|x_t)e^{\tilde{\zeta}_{t,t-1}(x_t,x_{t-1})}}{\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[e^{\tilde{\zeta}_{t,t-1}(x_t,X_{t-1})}]}$$

where

$$\begin{aligned}
&\tilde{\zeta}_{t,t-1}(x_t,x_{t-1}) \\
&:= \sqrt{\alpha_t}\Delta_t(x_t)^\intercal(x_{t-1}-m_t(x_t)) + (\nabla\log\tilde{q}_{t-1}(m_t(x_t)) - \sqrt{\alpha_t}\nabla\log\tilde{q}_t(x_t))^\intercal(x_{t-1}-m_t(x_t)) \\
&+ \sum_{p=2}^{\infty}T_p(\log\tilde{q}_{t-1}, x_{t-1}, m_t(x_t)).
\end{aligned}$$

Here we define the p-th order term in the Taylor expansion of $f(x)$ around μ as

$$T_p(f,x,\mu) := \frac{1}{p!}\sum_{\gamma\in\mathbb{N}^d:\sum_i\gamma^i=p}\partial_{\boldsymbol{a}}^p f(\mu)\prod_{i=1}^{d}(x^i-\mu^i)^{\gamma^i}$$

where $\boldsymbol{a}\in[d]^p$ are the indices of differentiation in which the multiplicity of $i\in[d]$ is γ^i.

Proof. See Appendix H.1. □

We abbreviate $\tilde{\zeta}_{t,t-1} = \tilde{\zeta}_{t,t-1}(x_t,x_{t-1})$. Given the expression of $\tilde{\zeta}_{t,t-1}$, the conditional reverse-step error can be upper-bounded for any fixed x_t as

$$\begin{aligned}
&\mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}}\left[\log\frac{\tilde{q}_{t-1|t}(X_{t-1}|x_t)}{p_{t-1|t}(X_{t-1}|x_t)}\right] \\
&= \mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}}\left[\tilde{\zeta}_{t,t-1} - \log\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[e^{\tilde{\zeta}_{t,t-1}}]\right]
\end{aligned}$$

$$\overset{(i)}{\leq} \mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}}\left[\tilde{\zeta}_{t,t-1}\right] + \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\left[-\log e^{\tilde{\zeta}_{t,t-1}}\right]$$
$$= \mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}}[\tilde{\zeta}_{t,t-1}] - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[\tilde{\zeta}_{t,t-1}] \tag{13}$$

where in (i) we use Jensen's inequality and note that $-\log(\cdot)$ is convex. Thus, from (12), we have an upper bound as

$$\mathrm{KL}(\tilde{Q}_0\|\widehat{P}_0) \leq \underbrace{\mathbb{E}_{X_T\sim\tilde{Q}_T}\left[\log\frac{\tilde{q}_T(X_T)}{\widehat{p}_T(X_T)}\right]}_{\text{Term 1: initialization error}} + \underbrace{\sum_{t=1}^{T}\mathbb{E}_{X_t,X_{t-1}\sim\tilde{Q}_{t,t-1}}\left[\log\frac{p_{t-1|t}(X_{t-1}|X_t)}{\widehat{p}_{t-1|t}(X_{t-1}|X_t)}\right]}_{\text{Term 2: estimation error}}$$
$$+ \underbrace{\sum_{t=1}^{T}\mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}}[\tilde{\zeta}_{t,t-1}] - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[\tilde{\zeta}_{t,t-1}]}_{\text{Term 3: reverse-step error}}.$$

Here, using (Liang et al., 2024, Lemma 3), the initialization error can be upper-bounded as, when $T\to\infty$,

$$\mathbb{E}_{X_T\sim\tilde{Q}_T}\left[\log\frac{\tilde{q}_T(X_T)}{\widehat{p}_T(X_T)}\right] \leq \frac{1}{2}\mathbb{E}_{X_0\sim\tilde{Q}_0}\|X_0\|^2\bar{\alpha}_T + O\left(\bar{\alpha}_T^2\right).$$

This implies that, under Assumption 1 and if $c>1$,

$$\mathbb{E}_{X_T\sim\tilde{Q}_T}\left[\log\frac{\tilde{q}_T(X_T)}{p_T(X_T)}\right] = o(T^{-1}).$$

Also, under Assumption 3, the higher-order Taylor polynomials enjoy exponential rate of decay in expectation, which is contained in powers of $(1-\alpha_t)$. Thus, we are allowed to exchange the limit (of Taylor expansion) and the expectation operators (cf. (Liang et al., 2024, Lemma 11)).

Now, we upper-bound the estimation error and reverse-step error under score mismatch separately.

F.1 Step 1: Bounding estimation error under mismatch

The following lemma provides an upper bound for the estimation error under score mismatch.

Lemma 2. *Under Assumptions 2 and 4, with the α_t satisfying Definition 1, we have*

$$\sum_{t=1}^{T}\mathbb{E}_{X_t,X_{t-1}\sim\tilde{Q}_{t,t-1}}\left[\log\frac{p_{t-1|t}(X_{t-1}|X_t)}{\widehat{p}_{t-1|t}(X_{t-1}|X_t)}\right] \lesssim \max_{t\geq 1}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\Delta_t(X_t)\|^2}(\log T)\varepsilon + (\log T)\varepsilon^2.$$

Proof. See Appendix H.2. □

F.2 Step 2: Decomposing reverse-step error under mismatch

Now, we decompose $\tilde{\zeta}_{t,t-1}(x_t,x_{t-1}) = \tilde{\zeta}_{\text{mis}} + \tilde{\zeta}_{\text{van}}$ where

$$\tilde{\zeta}_{\text{mis}} := \sqrt{\alpha_t}\Delta_t(x_t)^\intercal(x_{t-1}-m_t(x_t)),$$
$$\tilde{\zeta}_{\text{van}} := (\nabla\log\tilde{q}_{t-1}(m_t(x_t)) - \sqrt{\alpha_t}\nabla\log\tilde{q}_t(x_t))^\intercal(x_{t-1}-m_t(x_t)) + \sum_{p=2}^{\infty}T_p(\log\tilde{q}_{t-1},x_{t-1},m_t(x_t)). \tag{14}$$

Here $\tilde{\zeta}_{\text{van}}$ is the same tilting factor without score bias (cf. Liang et al. (2024)). Also, define an auxiliary conditional probability $\tilde{P}_{t-1|t}$ such that

$$\tilde{P}_{t-1|t} := \mathcal{N}\left(m_t(x_t), \frac{1-\alpha_t}{\alpha_t}I_d\right),$$

which corresponds to the oracle reverse process that knows the true scores of the perturbed target distributions. Thus, we can decompose the expected value of (13) in the following way:

$$\mathbb{E}_{X_t\sim\tilde{Q}_t}\left(\mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}} - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\right)[\tilde{\zeta}_{t,t-1}]$$

$$= \underbrace{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left(\mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}} - \mathbb{E}_{X_{t-1}\sim\tilde{P}_{t-1|t}}\right)[\tilde{\zeta}_{t,t-1}]}_{\mathcal{W}_{\text{oracle, rev-step}}}$$
$$+ \underbrace{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left(\mathbb{E}_{X_{t-1}\sim\tilde{P}_{t-1|t}} - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\right)[\tilde{\zeta}_{\text{mis}}]}_{\mathcal{W}_{\text{bias, rev-step}}}$$
$$+ \underbrace{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left(\mathbb{E}_{X_{t-1}\sim\tilde{P}_{t-1|t}} - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\right)[\tilde{\zeta}_{\text{van}}]}_{\mathcal{W}_{\text{vanish, rev-step}}}. \tag{15}$$

F.3 Step 3: Bounding $\mathcal{W}_{\text{oracle, rev-step}}$ and $\mathcal{W}_{\text{bias, rev-step}}$

Among the terms above, (Liang et al., 2024, Theorem 1) shows that, under Assumption 3 and using the α_t in Definition 1,

$$\mathcal{W}_{\text{oracle, rev-step}} \lesssim \sum_{t=1}^{T}(1-\alpha_t)^2\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\mathrm{Tr}\Big(\nabla^2\log\tilde{q}_{t-1}(m_t(X_t))\nabla^2\log\tilde{q}_t(X_t)\Big)\right]. \tag{16}$$

Also, for $\mathcal{W}_{\text{bias, rev-step}}$, since direct calculation yields

$$\mathbb{E}_{X_{t-1}\sim\tilde{P}_{t-1|t}}[\tilde{\zeta}_{\text{mis}}(x_t, X_{t-1})] = \mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}}[\tilde{\zeta}_{\text{mis}}(x_t, X_{t-1})] = 0,$$
$$\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[\tilde{\zeta}_{\text{mis}}(x_t, X_{t-1})] = -(1-\alpha_t)\left\|\Delta_t(x_t)\right\|^2,$$

we have

$$\mathcal{W}_{\text{bias, rev-step}} = \mathbb{E}_{X_t\sim\tilde{Q}_t}\left(\mathbb{E}_{X_{t-1}\sim\tilde{P}_{t-1|t}} - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\right)[\tilde{\zeta}_{\text{mis}}(X_t, X_{t-1})]$$
$$= (1-\alpha_t)\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2. \tag{17}$$

F.4 Step 4: Bounding $\mathcal{W}_{\text{vanish, rev-step}}$

Next, for $\mathcal{W}_{\text{vanish, rev-step}}$, we first note that $\tilde{P}_{t-1|t}$ is conditional Gaussian. Thus, under Assumption 3, we are able to exchange the limit (from Taylor series) and the expectation due to Gaussian-like moments (cf. (Liang et al., 2024, Lemma 11)), which gives us

$$\mathbb{E}_{X_{t-1}\sim\tilde{P}_{t-1|t}}[\tilde{\zeta}_{\text{van}}(x_t, X_{t-1})]$$
$$= \frac{1-\alpha_t}{2\alpha_t}\mathrm{Tr}(\nabla^2\log\tilde{q}_{t-1}(m_t(x_t))) + \sum_{p=4}^{\infty}\mathbb{E}_{X_{t-1}\sim\tilde{P}_{t-1|t}}\left[T_p(\log\tilde{q}_{t-1}, X_{t-1}, m_t(x_t))\right].$$

Here the expected value at $p=3$ is zero because all odd-order centralized moments of Gaussian vanish.

Now it remains to characterize the expectation of $\tilde{\zeta}_{\text{van}}(x_t, x_{t-1})$ under $P_{t-1|t}$. To this end, we introduce the following notation.

Definition 2 (Big-O in $\mathcal{L}^p$ space)**.** For a random variable Z_T, we say that $Z_T(x) = O_{\mathcal{L}^p(Q)}(1)$ if $(\mathbb{E}_{X\sim Q}\left|Z_T(X)\right|^p)^{1/p} = O(1)$ for all $p\geq 1$ as $T\to\infty$. Define $\tilde{O}_{\mathcal{L}^p(Q)}$ likewise.

One property is that if $Z_T(x) = O_{\mathcal{L}^p(Q)}(1)$ then $\mathbb{E}_{X\sim Q}\left|Z_T(X)\right| = O(1)$. Another property is that if $Z_1 = O_{\mathcal{L}^p(Q)}(a_T)$ and $Z_2 = O_{\mathcal{L}^p(Q)}(b_T)$ for all $p\geq 1$, applying Cauchy-Schwartz inequality we get, for all $p\geq 1$,

$$(\mathbb{E}_{X\sim Q}\left|Z_1Z_2\right|^p)^{1/p} \leq \left(\mathbb{E}_{X\sim Q}Z_1^{2p}\mathbb{E}_{X\sim Q}Z_2^{2p}\right)^{1/(2p)} = O(a_Tb_T),$$

which implies that $O_{\mathcal{L}^p(Q)}(a_T)O_{\mathcal{L}^p(Q)}(b_T) = O_{\mathcal{L}^p(Q)}(a_Tb_T)$. Now, with this notation, the first lines of Assumption 3 can be equivalently written as

$$(1-\alpha_t)^m\left|\partial_{\boldsymbol{a}}^k\log q_t(X_t)\right| = O_{\mathcal{L}^p(\tilde{Q}_t)}\left((1-\alpha_t)^m\right),\ \forall p\geq 1,$$
$$(1-\alpha_t)^m\left|\partial_{\boldsymbol{a}}^k\log q_{t-1}(m_t(X_t))\right| = O_{\mathcal{L}^p(\tilde{Q}_t)}\left((1-\alpha_t)^m\right),\ \forall p\geq 1.$$

Also, Assumption 4 can be equivalently written as

$$(1-\alpha_t)^m \left\|\Delta_{t,y}(X_t)\right\| = O_{\mathcal{L}^p(\tilde{Q}_t)}(\bar{\alpha}_t(1-\alpha_t)^m),\ \forall p \geq 1.$$

With these notations, the following lemma characterizes the expectation of $\tilde{\zeta}_{\text{van}}(x_t, x_{t-1})$ under $P_{t-1|t}$, which involves non-centralized Gaussian moments.

Lemma 3. *As long as $\tilde{q}_{t-1}$ is defined, with the definition of $\tilde{\zeta}_{van}$ in* (14)*, under Assumptions 3 and 4, we have $\forall \ell \geq 1$,*

$$\begin{aligned}
&\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[\tilde{\zeta}_{van}(x_t, X_{t-1})] \\
&= -\frac{1-\alpha_t}{\sqrt{\alpha_t}}(\nabla \log \tilde{q}_{t-1}(m_t(x_t)) - \sqrt{\alpha_t}\nabla \log \tilde{q}_t(x_t))^\intercal \Delta_t(x_t) \\
&\quad + \frac{1-\alpha_t}{2\alpha_t}\mathrm{Tr}(\nabla^2 \log \tilde{q}_{t-1}(m_t)) + \frac{(1-\alpha_t)^2}{2\alpha_t}\Delta_t(x_t)^\intercal \nabla^2 \log \tilde{q}_{t-1}(m_t)\Delta_t(x_t) \\
&\quad - \frac{1}{3!}\left(\frac{(1-\alpha_t)^2}{{\alpha_t}^{3/2}}\right)\left(3\sum_{i=1}^{d}\partial^3_{iii}\log \tilde{q}_{t-1}(m_t)\Delta_t^i + \sum_{\substack{i,j=1\\ i\neq j}}^{d}\partial^3_{iij}\log \tilde{q}_{t-1}(m_t)\Delta_t^j\right) \\
&\quad + \sum_{p=4}^{\infty}\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\left[T_p(\log \tilde{q}_{t-1}, X_{t-1}, m_t(x_t))\right] + O_{\mathcal{L}^\ell(\tilde{Q}_t)}\left((1-\alpha_t)^3\right).
\end{aligned}$$

Proof. See Appendix H.3. □

The following lemma provides the rate of decay of the difference in expectation of all Taylor polynomials with order $p \geq 4$.

Lemma 4. *As long as $\tilde{q}_{t-1}$ is defined, under Assumptions 3 and 4, we have, $\forall p \geq 4, \ell \geq 1$,*

$$\left(\mathbb{E}_{X_{t-1}\sim \tilde{P}_{t-1|t}} - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\right)\left[T_p(\log \tilde{q}_{t-1}, X_{t-1}, m_t(X_t))\right] = O_{\mathcal{L}^\ell(\tilde{Q}_t)}\left((1-\alpha_t)^3\right).$$

Proof. See Appendix H.4. □

Thus, with the help of Lemmas 3 and 4, we can identify the dominating terms in $\mathcal{W}_{\text{vanish, rev-step}}$ when $1-\alpha_t$ is small. The dominating term is

$$\begin{aligned}
\mathcal{W}_{\text{vanish, rev-step}} &= \mathbb{E}_{X_t\sim\tilde{Q}_t}\left(\mathbb{E}_{X_{t-1}\sim\tilde{P}_{t-1|t}} - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\right)\left[\tilde{\zeta}_{\text{van}}(X_t, X_{t-1})\right] \\
&= \frac{1-\alpha_t}{\sqrt{\alpha_t}}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[(\nabla \log \tilde{q}_{t-1}(m_t(X_t)) - \sqrt{\alpha_t}\nabla \log \tilde{q}_t(X_t))^\intercal \Delta_t(X_t)\right] \\
&\quad - \frac{(1-\alpha_t)^2}{2\alpha_t}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\Delta_t(X_t)^\intercal \nabla^2 \log \tilde{q}_{t-1}(m_t(X_t))\Delta_t(X_t)\right] \\
&\quad + \frac{1}{3!}\left(\frac{(1-\alpha_t)^2}{{\alpha_t}^{3/2}}\right)\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[3\sum_{i=1}^{d}\partial^3_{iii}\log \tilde{q}_{t-1}(m_t(X_t))\Delta_t(X_t)^i\right. \\
&\qquad \left. + \sum_{\substack{i,j=1\\ i\neq j}}^{d}\partial^3_{iij}\log \tilde{q}_{t-1}(m_t(X_t))\Delta_t(X_t)^j\right] \\
&\quad + O((1-\alpha_t)^3).
\end{aligned} \tag{18}$$

Therefore, with the decomposition in (15) in mind, an upper bound on the reverse-step error is achieved by summing up (16), (17) and (18). The proof of Theorem 1 is now complete.

F.5 COROLLARY 1 AND ITS PROOF

Below we state and prove a corollary of Theorem 1 when $\tilde{q}_0$ does not exist. By (Liang et al., 2024, Lemma 6), $\tilde{q}_1$ always exists, which provides us with the following convergence result with early-stopping.

Corollary 1. *Suppose that Assumptions 1 to 4 are satisfied. Then, suppose that the α_t satisfies Definition 1 at $t \geq 2$, the distribution $\widehat{P}_1$ from the discrete-time DDPM under score bias satisfies*

$$
\begin{aligned}
&\mathrm{KL}(\tilde{Q}_1 \| \widehat{P}_1) \\
&\lesssim \sum_{t=2}^{T} (1-\alpha_t) \mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \Delta_t(X_t) \right\|^2 \\
&+ \sum_{t=2}^{T} \frac{1-\alpha_t}{\sqrt{\alpha_t}} \mathbb{E}_{X_t \sim \tilde{Q}_t} \left[(\nabla \log \tilde{q}_{t-1}(m_t(X_t)) - \sqrt{\alpha_t} \nabla \log \tilde{q}_t(X_t))^\intercal \Delta_t(X_t) \right] \\
&+ \sum_{t=2}^{T} \frac{(1-\alpha_t)^2}{2\alpha_t} \mathbb{E}_{X_t \sim \tilde{Q}_t} \left[\mathrm{Tr}\Big(\nabla^2 \log \tilde{q}_{t-1}(m_t(X_t)) \left(\nabla^2 \log \tilde{q}_t(X_t) - \Delta_t(X_t) \Delta_t(X_t)^\intercal \right) \Big) \right] \\
&+ \sum_{t=2}^{T} \frac{(1-\alpha_t)^2}{3! {\alpha_t}^{3/2}} \mathbb{E}_{X_t \sim \tilde{Q}_t} \left[3 \sum_{i=1}^{d} \partial^3_{iii} \log \tilde{q}_{t-1}(m_t(X_t)) \Delta_t(X_t)^i + \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \partial^3_{iij} \log \tilde{q}_{t-1}(m_t(X_t)) \Delta_t(X_t)^j \right] \\
&+ \max_{t \geq 2} \sqrt{\mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \Delta_t(X_t) \right\|^2} (\log T) \varepsilon + (\log T) \varepsilon^2,
\end{aligned}
$$

where $\mathrm{W}_2(\tilde{Q}_1, \tilde{Q}_0)^2 \lesssim (1-\alpha_1) d$.

Proof. The result directly follows with the same arguments as in the proof of Theorem 1. The only difference is the guarantee under the Wasserstein distance, which can be obtained using (Liang et al., 2024, Lemma 12). □

G PROOF OF THEOREM 2

We first recall some of the properties of the noise schedule in (8). By Lemma 6, the noise schedule in (8) satisfies that $\frac{1-\alpha_t}{(1-\bar{\alpha}_{t-1})^p} \lesssim \frac{\log T \log(1/\delta)}{\delta^{p-1} T}$ for all $p \geq 1$ while $\bar{\alpha}_T = o(T^{-1})$, and thus such α_t satisfies Definition 1 when $t \geq 2$. Further, with the α_t in (8), (Liang et al., 2024, Lemmas 15 and 17) show that for any Q_0 with finite variance under early-stopping, $\forall p, \ell \geq 1$,

$$
\begin{aligned}
\mathbb{E}_{X_t \sim \tilde{Q}_t} \left| \partial^p_{\boldsymbol{a}} \log \tilde{q}_t(X_t) \right|^\ell &= O\left(\frac{1}{(1-\bar{\alpha}_t)^{p\ell/2}} \right), \\
\mathbb{E}_{X_t \sim \tilde{Q}_t} \left| \partial^p_{\boldsymbol{a}} \log \tilde{q}_{t-1}(m_t(X_t)) \right|^\ell &= O\left(\frac{1}{(1-\bar{\alpha}_{t-1})^{p\ell/2}} \right).
\end{aligned}
$$

Thus, using Lemma 6, Assumption 3 is satisfied (since δ is constant). In the following, we further verify the last relationship in Assumption 3 holds.

Therefore, since Assumptions 1, 2 and 4 have been satisfied, we can invoke Corollary 1 and get $\mathrm{KL}(\tilde{Q}_1 \| \widehat{P}_1) \lesssim \mathcal{W}_{\text{oracle}} + \mathcal{W}_{\text{bias}} + \mathcal{W}_{\text{vanish}}$. Now, we investigate the dimensional dependence for each term of the upper bound in Corollary 1.

To start, from (Liang et al., 2024, Theorem 3), for any $\tilde{Q}_0$ having finite variance, with the α_t in (8), we have

$$
\mathcal{W}_{\text{oracle}} \lesssim \frac{d^2 \log^2(1/\delta) (\log T)^2}{T} + (\log T) \varepsilon^2.
$$

Also, since by assumption $\mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \Delta_t(X_t) \right\|^2 \lesssim \frac{\bar{\alpha}_t}{(1-\bar{\alpha}_t)^r} d^\gamma$, with the α_t in (8), we have from Lemma 7 that when $\delta \ll 1$,

$$
\mathcal{W}_{\text{bias}} \lesssim \frac{d^\gamma}{\delta^r} \left(1 - \frac{2 \log(1/\delta) \log T}{T} \right).
$$

Now we investigate each term in $\mathcal{W}_{\text{vanish}}$. The following lemma is useful to determine the rate of difference of the first-order Taylor polynomials.

Lemma 5. *When $\mathbb{E}_{X_0\sim\tilde{Q}_0}\|X_0\|^6 \lesssim d^3$, with the α_t in (8), we have*

$$(1-\alpha_t)\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\nabla\log\tilde{q}_{t-1}(m_t(X_t)) - \sqrt{\alpha_t}\nabla\log\tilde{q}_t(X_t)\right\|^2} \lesssim \frac{d^{3/2}(1-\alpha_t)^2}{(1-\bar{\alpha}_{t-1})^3}.$$

As a result, Assumption 3 holds.

Proof. See Appendix H.5. □

In other words, combining Lemma 5 and Lemma 6, we have

$$(1-\alpha_t)\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\nabla\log\tilde{q}_{t-1}(m_t(X_t)) - \sqrt{\alpha_t}\nabla\log\tilde{q}_t(X_t)\right\|^2} = \tilde{O}\left(\frac{1}{T^2}\right). \tag{19}$$

Now, by Cauchy-Schwartz inequality and Lemma 5,

$$\begin{aligned}
&\sum_{t=2}^{T}\frac{1-\alpha_t}{\sqrt{\alpha_t}}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[(\nabla\log\tilde{q}_{t-1}(m_t(X_t)) - \sqrt{\alpha_t}\nabla\log\tilde{q}_t(X_t))^\intercal\Delta_t(X_t)\right]\\
&\le \sum_{t=2}^{T}\frac{1-\alpha_t}{\sqrt{\alpha_t}}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2}\times\\
&\qquad\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\nabla\log\tilde{q}_{t-1}(m_t(X_t)) - \sqrt{\alpha_t}\nabla\log\tilde{q}_t(X_t)\right\|^2}\\
&\lesssim \frac{d^{\gamma/2}}{(1-\bar{\alpha}_t)^{r/2}}\times\frac{d^{3/2}(1-\alpha_t)^2}{(1-\bar{\alpha}_{t-1})^3}\\
&\lesssim \sum_{t=2}^{T}\frac{d^{\frac{3+\gamma}{2}}\log(1/\delta)^2(\log T)^2}{\delta^{1+r/2}T^2}\\
&\le \frac{d^{\frac{3+\gamma}{2}}\log(1/\delta)^2(\log T)^2}{\delta^{1+r/2}T}.
\end{aligned}$$

To proceed for higher orders of Taylor polynomials, we first note that from (Liang et al., 2024, Section G.2), the second and third derivatives of $\log\tilde{q}_t$ are

$$\begin{aligned}
\nabla^2\log\tilde{q}_t(x) &= -\frac{1}{1-\bar{\alpha}_t}I_d + \frac{1}{(1-\bar{\alpha}_t)^2}\Bigg(\mathbb{E}_{X_0\sim\tilde{Q}_{0|t}(\cdot|x)}\left[(x-\sqrt{\bar{\alpha}_t}X_0)(x-\sqrt{\bar{\alpha}_t}X_0)^\intercal\right]\\
&\qquad - \left(\mathbb{E}_{X_0\sim\tilde{Q}_{0|t}(\cdot|x)}\left[x-\sqrt{\bar{\alpha}_t}X_0\right]\right)\left(\mathbb{E}_{X_0\sim\tilde{Q}_{0|t}(\cdot|x)}\left[x-\sqrt{\bar{\alpha}_t}X_0\right]\right)^\intercal\Bigg)\\
\partial^3_{ijk}\log\tilde{q}_t(x) &= -\int z^i z^j z^k \mathrm{d}\tilde{Q}_{0|t}(x_0|x)\\
&\qquad + \sum_{\substack{a_1=i,j,k\\ a_2<a_3,\ a_2,a_3\neq a_1}}\int z^{a_1}\mathrm{d}\tilde{Q}_{0|t}(x_0|x)\int z^{a_2}z^{a_3}\mathrm{d}\tilde{Q}_{0|t}(x_0|x)\\
&\qquad - 2\int z^i\mathrm{d}\tilde{Q}_{0|t}(x_0|x)\int z^j\mathrm{d}\tilde{Q}_{0|t}(x_0|x)\int z^k\mathrm{d}\tilde{Q}_{0|t}(x_0|x)
\end{aligned}$$

where $z := \frac{x-\sqrt{\bar{\alpha}_t}x_0}{1-\bar{\alpha}_t}$. Thus, in order to provide an upper bound on the expected norm of the second-order derivative of $\log\tilde{q}_{t-1}(m_t)$, we can calculate

$$\begin{aligned}
&\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\mathbb{E}_{X_0\sim\tilde{Q}_{0|t-1}(\cdot|m_t)}\left[(m_t-\sqrt{\bar{\alpha}_{t-1}}X_0)(m_t-\sqrt{\bar{\alpha}_{t-1}}X_0)^\intercal\right]\right\|_F^2\\
&\le \mathbb{E}_{X_t\sim\tilde{Q}_t,X_0\sim\tilde{Q}_{0|t-1}(\cdot|m_t)}\left\|(m_t-\sqrt{\bar{\alpha}_{t-1}}X_0)(m_t-\sqrt{\bar{\alpha}_{t-1}}X_0)^\intercal\right\|_F^2\\
&= \mathbb{E}_{X_t\sim\tilde{Q}_t,X_0\sim\tilde{Q}_{0|t-1}(\cdot|m_t)}\left\|m_t-\sqrt{\bar{\alpha}_{t-1}}X_0\right\|^4
\end{aligned}$$

$$\overset{(i)}{\lesssim} d^2(1-\bar{\alpha}_{t-1})^2,$$

and

$$\begin{aligned}
&\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\left(\mathbb{E}_{X_0\sim\tilde{Q}_{0|t-1}}\left[m_t-\sqrt{\bar{\alpha}_{t-1}}X_0\right]\right)\left(\mathbb{E}_{X_0\sim\tilde{Q}_{0|t-1}}\left[m_t-\sqrt{\bar{\alpha}_{t-1}}X_0\right]\right)^\intercal\right\|_F^2\\
&=\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\mathbb{E}_{X_0\sim\tilde{Q}_{0|t-1}(\cdot|m_t)}\left[m_t-\sqrt{\bar{\alpha}_{t-1}}X_0\right]\right\|^4\\
&\le\mathbb{E}_{X_t\sim\tilde{Q}_t,X_0\sim\tilde{Q}_{0|t-1}(\cdot|m_t)}\left\|m_t-\sqrt{\bar{\alpha}_{t-1}}X_0\right\|^4\\
&\overset{(ii)}{\lesssim} d^2(1-\bar{\alpha}_{t-1})^2,
\end{aligned}$$

where both (i) and (ii) follow from (Liang et al., 2024, Lemma 16). Thus,

$$\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\nabla^2\log\tilde{q}_{t-1}(m_t(X_t))\right\|_F^2\lesssim\frac{1}{(1-\bar{\alpha}_{t-1})^2}d^2.$$

For third-order derivatives, we can similarly use (Liang et al., 2024, Lemma 16) and get (cf. (Liang et al., 2024, Section G.2))

$$\begin{aligned}
\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\sum_{i=1}^d(\partial^3_{iii}\log\tilde{q}_{t-1}(m_t(X_t)))^2\right]&\lesssim\frac{1}{(1-\bar{\alpha}_{t-1})^3}\sum_{i=1}^d\mathbb{E}(Z^i)^6\lesssim\frac{d}{(1-\bar{\alpha}_{t-1})^3},\\
\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\sum_{i,j=1}^d(\partial^3_{iij}\log\tilde{q}_{t-1}(m_t(X_t)))^2\right]&\lesssim\frac{1}{(1-\bar{\alpha}_{t-1})^3}\sum_{i,j=1}^d\left(\mathbb{E}(Z^i)^6\right)^{2/3}\left(\mathbb{E}(Z^j)^6\right)^{1/3}\\
&\lesssim\frac{d^2}{(1-\bar{\alpha}_{t-1})^3}.
\end{aligned}$$

Here we denote $Z\sim\mathcal{N}(0,I_d)$, and note that $\mathbb{E}(Z^i)^6,\mathbb{E}(Z^j)^6\lesssim 1$.

Therefore, by Cauchy-Schwartz inequality,

$$\begin{aligned}
&\sum_{t=2}^T\frac{(1-\alpha_t)^2}{2\alpha_t}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\Delta_t(X_t)^\intercal\nabla^2\log\tilde{q}_{t-1}(m_t(X_t))\Delta_t(X_t)\right]\\
&\le\sum_{t=2}^T\frac{(1-\alpha_t)^2}{2\alpha_t}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^4}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\nabla^2\log\tilde{q}_{t-1}(m_t(X_t))\right\|^2}\\
&\lesssim\sum_{t=2}^T\frac{(1-\alpha_t)^2}{2\alpha_t}\frac{d^\gamma}{(1-\bar{\alpha}_t)^r}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\nabla^2\log\tilde{q}_{t-1}(m_t(X_t))\right\|_F^2}\\
&\lesssim\sum_{t=2}^T\frac{(1-\alpha_t)^2}{2\alpha_t}\frac{d^\gamma}{(1-\bar{\alpha}_t)^r}\sqrt{\frac{d^2}{(1-\bar{\alpha}_{t-1})^2}}\\
&\lesssim\frac{d^{1+\gamma}\log(1/\delta)^2(\log T)^2}{\delta^{r-1}T},
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{t=2}^T\frac{(1-\alpha_t)^2}{3!{\alpha_t}^{3/2}}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[3\sum_{i=1}^d\partial^3_{iii}\log\tilde{q}_{t-1}(m_t(X_t))\Delta_t(X_t)^i\right]\\
&\le\sum_{t=2}^T\frac{3(1-\alpha_t)^2}{3!{\alpha_t}^{3/2}}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\sum_{i=1}^d(\partial^3_{iii}\log\tilde{q}_{t-1}(m_t(X_t)))^2}\\
&\lesssim\sum_{t=2}^T(1-\alpha_t)^2\frac{d^{\gamma/2}}{(1-\bar{\alpha}_t)^{r/2}}\sqrt{\frac{d}{(1-\bar{\alpha}_{t-1})^3}}
\end{aligned}$$

$$\lesssim \frac{d^{\frac{1+\gamma}{2}}\log(1/\delta)^2(\log T)^2}{\delta^{\frac{r-1}{2}}T},$$

and, with M being a matrix such that $M^{ij}(x) := \partial^3_{iij}\log\tilde{q}_{t-1}(m_t(x))$,

$$\begin{aligned}
&\sum_{t=2}^{T}\frac{(1-\alpha_t)^2}{3!{\alpha_t}^{3/2}}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\sum_{\substack{i,j=1\\ i\neq j}}^{d}\partial^3_{iij}\log\tilde{q}_{t-1}(m_t(X_t))\Delta_t(X_t)^j\right]\\
&\le \sum_{t=2}^{T}\frac{(1-\alpha_t)^2}{3!{\alpha_t}^{3/2}}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|M(X_t)\Delta_t(X_t)\right\|_1\\
&\le \sum_{t=2}^{T}\frac{(1-\alpha_t)^2}{3!{\alpha_t}^{3/2}}\sqrt{d}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|M(X_t)\right\|^2}\\
&\le \sum_{t=2}^{T}\frac{(1-\alpha_t)^2}{3!{\alpha_t}^{3/2}}\sqrt{d}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\sum_{i,j=1}^{d}(\partial^3_{iij}\log\tilde{q}_{t-1}(m_t(X_t)))^2\right]}\\
&\lesssim \sum_{t=2}^{T}(1-\alpha_t)^2\frac{d^{(1+\gamma)/2}}{(1-\bar{\alpha}_t)^{r/2}}\sqrt{\frac{d^2}{(1-\bar{\alpha}_{t-1})^3}}\\
&\lesssim \frac{d^{\frac{3+\gamma}{2}}\log(1/\delta)^2(\log T)^2}{\delta^{\frac{r-1}{2}}T},
\end{aligned}$$

and

$$\max_{t\ge 2}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2}(\log T)\varepsilon \lesssim \frac{d^{\gamma/2}}{\delta^{r/2}}(\log T)\varepsilon.$$

Therefore, combining all the above, we get

$$\begin{aligned}
\mathrm{KL}(\tilde{Q}_1\|\widehat{P}_1) \lesssim\ & d^{\gamma}\delta^{-r}\left(1-\frac{2\log(1/\delta)\log T}{T}\right)\\
&+\max\{d^{(3+\gamma)/2}\delta^{-\frac{r+2}{2}}, d^{1+\gamma}\delta^{-(r-1)}\}\frac{(\log T)^2}{T}\\
&+d^{\gamma/2}\delta^{-r/2}(\log T)\varepsilon.
\end{aligned}$$

H Auxiliary Lemmas and Proofs in Section 3

H.1 Proof of Lemma 1

We remind readers that throughout this proof x_t is fixed. For brevity write $m_t = m_t(x_t)$, $\mu_t = \mu_t(x_t)$, and $\Delta_t(x_t) = \Delta_t$. Recall that $m_t = x_t/\sqrt{\alpha_t} + (1-\alpha_t)/\sqrt{\alpha_t}\nabla\log\tilde{q}_t(x_t)$. By Bayes' rule, we have

$$\begin{aligned}
&\tilde{q}_{t-1|t}(x_{t-1}|x_t)\\
&= \frac{\tilde{q}_{t|t-1}(x_t|x_{t-1})\tilde{q}_{t-1}(x_{t-1})}{\tilde{q}_t(x_t)}\\
&\propto \tilde{q}_{t-1}(x_{t-1})\tilde{q}_{t|t-1}(x_t|x_{t-1})\\
&\overset{(i)}{\propto} \tilde{q}_{t-1}(x_{t-1})\exp\left(-\frac{\left\|x_t-\sqrt{\alpha_t}x_{t-1}\right\|^2}{2(1-\alpha_t)}\right)\\
&\propto \tilde{q}_{t-1}(x_{t-1})p_{t-1|t}(x_{t-1}|x_t)\exp\left(\frac{\left\|x_{t-1}-\mu_t\right\|^2-\left\|x_{t-1}-x_t/\sqrt{\alpha_t}\right\|^2}{2(1-\alpha_t)/\alpha_t}\right)\\
&= \tilde{q}_{t-1}(x_{t-1})p_{t-1|t}(x_{t-1}|x_t)\exp\left(\frac{\left\|x_{t-1}-x_t/\sqrt{\alpha_t}+x_t/\sqrt{\alpha_t}-\mu_t\right\|^2-\left\|x_{t-1}-x_t/\sqrt{\alpha_t}\right\|^2}{2(1-\alpha_t)/\alpha_t}\right)
\end{aligned}$$

$$\propto \tilde{q}_{t-1}(x_{t-1})p_{t-1|t}(x_{t-1}|x_t)\exp\left(\frac{(x_{t-1}-x_t/\sqrt{\alpha_t})^\intercal(x_t/\sqrt{\alpha_t}-\mu_t)}{(1-\alpha_t)/\alpha_t}\right)$$

where (i) follows because the forward process is Markov and $\tilde{q}_{t|t-1}=q_{t|t-1}$. Here, the exponent is equal to

$$\begin{aligned}
&\frac{(x_{t-1}-x_t/\sqrt{\alpha_t})^\intercal(x_t/\sqrt{\alpha_t}-\mu_t)}{(1-\alpha_t)/\alpha_t}\\
&=\frac{(x_{t-1}-x_t/\sqrt{\alpha_t})^\intercal(m_t-\mu_t)}{(1-\alpha_t)/\alpha_t}-\frac{(x_{t-1}-x_t/\sqrt{\alpha_t})^\intercal((1-\alpha_t)/\sqrt{\alpha_t})\nabla\log\tilde{q}_t(x_t)}{(1-\alpha_t)/\alpha_t}\\
&=\sqrt{\alpha_t}\Delta_t^\intercal(x_{t-1}-x_t/\sqrt{\alpha_t})-\sqrt{\alpha_t}(x_{t-1}-x_t/\sqrt{\alpha_t})^\intercal\nabla\log\tilde{q}_t(x_t).
\end{aligned}$$

Thus,

$$\tilde{q}_{t-1|t}(x_{t-1}|x_t)\propto p_{t-1|t}(x_{t-1}|x_t)\exp\left(\tilde{\zeta}_{t,t-1}(x_t,x_{t-1})\right)$$

where

$$\tilde{\zeta}_{t,t-1}(x_t,x_{t-1})=\sqrt{\alpha_t}\Delta_t^\intercal(x_{t-1}-m_t)+\log\tilde{q}_{t-1}(x_{t-1})-\sqrt{\alpha_t}(x_{t-1}-m_t)^\intercal\nabla\log\tilde{q}_t(x_t).$$

Finally, since all partial derivatives of $\tilde{q}_{t-1}$ exists for any $t\geq 2$ (See (Liang et al., 2024, Lemma 6)), the Taylor expansion of $\log\tilde{q}_{t-1}$ around m_t gives the desirable result.

H.2 Proof of Lemma 2

For each $t=1,\ldots,T$,

$$\begin{aligned}
\log\frac{p_{t-1|t}(x_{t-1}|x_t)}{\widehat{p}_{t-1|t}(x_{t-1}|x_t)}&=\frac{\alpha_t}{2(1-\alpha_t)}\left(\|x_{t-1}-\widehat{\mu}_t(x_t)\|^2-\|x_{t-1}-\mu_t(x_t)\|^2\right)\\
&=\frac{\alpha_t}{(1-\alpha_t)}(x_{t-1}-\mu_t(x_t))^\intercal(\mu_t(x_t)-\widehat{\mu}_t(x_t))+\frac{\alpha_t}{2(1-\alpha_t)}\|\mu_t(x_t)-\widehat{\mu}_t(x_t)\|^2.
\end{aligned}$$

For the first term above,

$$\begin{aligned}
&\mathbb{E}_{X_t,X_{t-1}\sim\tilde{Q}_{t,t-1}}\left[(X_{t-1}-\mu_t(X_t))^\intercal(\mu_t(X_t)-\widehat{\mu}_t(X_t))\right]\\
&=\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[(m_t(X_t)-\mu_t(X_t))^\intercal(\mu_t(X_t)-\widehat{\mu}_t(X_t))\right]\\
&=\frac{1-\alpha_t}{\sqrt{\alpha_t}}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left[\Delta_t(X_t)^\intercal(\mu_t(X_t)-\widehat{\mu}_t(X_t))\right]\\
&\leq\frac{1-\alpha_t}{\sqrt{\alpha_t}}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\Delta_t(X_t)\|^2\,\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\|^2}.
\end{aligned}$$

Here we recall the definition of Δ_t from (2) where $m_t(x)-\mu_t(x)=\frac{1-\alpha_t}{\sqrt{\alpha_t}}\Delta_t(x)$. Thus,

$$\begin{aligned}
&\sum_{t=1}^T\mathbb{E}_{X_t,X_{t-1}\sim\tilde{Q}_{t,t-1}}\left[\log\frac{p_{t-1|t}(X_{t-1}|X_t)}{\widehat{p}_{t-1|t}(X_{t-1}|X_t)}\right]\\
&\lesssim\sum_{t=1}^T\left(\sqrt{\alpha_t}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\Delta_t(X_t)\|^2\,\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\|^2}+\frac{\alpha_t}{1-\alpha_t}\mathbb{E}_{X_t\sim Q_t}\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\|^2\right)\\
&=\sum_{t=1}^T(1-\alpha_t)\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\Delta_t(X_t)\|^2}\times\sqrt{\frac{\alpha_t}{(1-\alpha_t)^2}\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\|^2}\\
&\quad+\sum_{t=1}^T(1-\alpha_t)\frac{\alpha_t}{(1-\alpha_t)^2}\mathbb{E}_{X_t\sim Q_t}\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\|^2\\
&\overset{(i)}{\lesssim}\frac{\log T}{T}\sum_{t=1}^T\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\Delta_t(X_t)\|^2}\times\sqrt{\frac{\alpha_t}{(1-\alpha_t)^2}\mathbb{E}_{X_t\sim\tilde{Q}_t}\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\|^2}\\
&\quad+\frac{\log T}{T}\sum_{t=1}^T\frac{\alpha_t}{(1-\alpha_t)^2}\mathbb{E}_{X_t\sim Q_t}\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{t\geq 1}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2}\frac{\log T}{T}\sum_{t=1}^{T}\sqrt{\frac{\alpha_t}{(1-\alpha_t)^2}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\right\|^2} \\
&\quad+\frac{\log T}{T}\sum_{t=1}^{T}\frac{\alpha_t}{(1-\alpha_t)^2}\mathbb{E}_{X_t\sim Q_t}\left\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\right\|^2 \\
&\overset{(ii)}{\lesssim} \max_{t\geq 1}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2}(\log T)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{\alpha_t}{(1-\alpha_t)^2}\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\right\|^2} \\
&\quad+\frac{\log T}{T}\sum_{t=1}^{T}\frac{\alpha_t}{(1-\alpha_t)^2}\mathbb{E}_{X_t\sim Q_t}\left\|\mu_t(X_t)-\widehat{\mu}_t(X_t)\right\|^2 \\
&\overset{(iii)}{\lesssim} \max_{t\geq 1}\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\Delta_t(X_t)\right\|^2}(\log T)\varepsilon+(\log T)\varepsilon^2
\end{aligned}
$$

where (i) follows from Definition 1, (ii) follows from the fact that for any non-negative sequence a_t, $\frac{1}{T}\sum_{t=1}^{T}\sqrt{a_t}\leq\sqrt{\frac{1}{T}\sum_{t=1}^{T}a_t}$ by Jensen's inequality, and (iii) follows from Assumption 2. The proof is complete.

H.3 Proof of Lemma 3

Recall that $P_{t-1|t}=\mathcal{N}(\mu_t,\frac{1-\alpha_t}{\alpha_t}I_d)$, and thus $\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[X_{t-1}-m_t(x_t)]=-\frac{1-\alpha_t}{\sqrt{\alpha_t}}\Delta_t(x_t)$. Note that we can change the limit and the expectation under Assumption 3. Now, we can calculate that

$$
\begin{aligned}
&\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[\tilde{\zeta}_{\text{van}}(x_t,X_{t-1})] \\
&=-\frac{1-\alpha_t}{\sqrt{\alpha_t}}(\nabla\log\tilde{q}_{t-1}(m_t(x_t))-\sqrt{\alpha_t}\nabla\log\tilde{q}_t(x_t))^\intercal\Delta_t(x_t) \\
&+\sum_{p=2}^{\infty}\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\left[T_p(\log\tilde{q}_{t-1},X_{t-1},m_t(x_t))\right].
\end{aligned}
$$

Below we write $m_t=m_t(x_t)$. Since T_2 is in quadratic form, the expected value under $P_{t-1|t}$ for this term is

$$
\begin{aligned}
&\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\left[T_2(\log\tilde{q}_{t-1},X_{t-1},m_t)\right] \\
&\quad=\frac{1-\alpha_t}{2\alpha_t}\mathrm{Tr}(\nabla^2\log\tilde{q}_{t-1}(m_t))+\frac{(1-\alpha_t)^2}{2\alpha_t}\Delta_t(x_t)^\intercal\nabla^2\log\tilde{q}_{t-1}(m_t)\Delta_t(x_t).
\end{aligned}
$$

Recall the formula for Gaussian non-centralized third moment. If $Z\sim\mathcal{N}(\mu,\sigma^2)$, then $\mathbb{E}[Z^2]=\mu^2+\sigma^2$ and $\mathbb{E}[Z^3]=\mu^3+3\mu\sigma^2$. Thus, the expected value under $P_{t-1|t}$ for T_3 is

$$
\begin{aligned}
&\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}\left[T_3(\log\tilde{q}_{t-1},X_{t-1},m_t)\right] \\
&=\frac{1}{3!}\sum_{i=1}^{d}\partial^3_{iii}\log\tilde{q}_{t-1}(m_t)\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}(X^i_{t-1}-m^i_t)^3 \\
&\quad+\frac{1}{3!}\sum_{\substack{i,j=1\\ i\neq j}}^{d}\partial^3_{iij}\log\tilde{q}_{t-1}(m_t)\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}(X^i_{t-1}-m^i_t)^2(X^j_{t-1}-m^j_t) \\
&\quad+\frac{1}{3!}\sum_{\substack{i,j,k=1\\ i,j,k\text{ all differ}}}^{d}\partial^3_{ijk}\log\tilde{q}_{t-1}(m_t)\mathbb{E}_{X_{t-1}\sim P_{t-1|t}}(X^i_{t-1}-m^i_t)(X^j_{t-1}-m^j_t)(X^k_{t-1}-m^k_t) \\
&=\frac{1}{3!}\sum_{i=1}^{d}\partial^3_{iii}\log\tilde{q}_{t-1}(m_t)\left(\left(-\frac{1-\alpha_t}{\sqrt{\alpha_t}}\Delta^i_t\right)^3+3\left(-\frac{1-\alpha_t}{\sqrt{\alpha_t}}\Delta^i_t\right)\left(\frac{1-\alpha_t}{\alpha_t}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{3!} \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \partial^3_{iij} \log \tilde{q}_{t-1}(m_t) \left(\left(-\frac{1-\alpha_t}{\sqrt{\alpha_t}} \Delta_t^i \right)^2 + \left(\frac{1-\alpha_t}{\alpha_t} \right) \right) \left(-\frac{1-\alpha_t}{\sqrt{\alpha_t}} \Delta_t^j \right) \\
&+ \frac{1}{3!} \sum_{\substack{i,j,k=1 \\ i,j,k \text{ all differ}}}^{d} \partial^3_{ijk} \log \tilde{q}_{t-1}(m_t) \left(-\frac{1-\alpha_t}{\sqrt{\alpha_t}} \Delta_t^i \right) \left(-\frac{1-\alpha_t}{\sqrt{\alpha_t}} \Delta_t^j \right) \left(-\frac{1-\alpha_t}{\sqrt{\alpha_t}} \Delta_t^k \right) \\
&= \frac{1}{3!} \left(-\frac{(1-\alpha_t)^2}{{\alpha_t}^{3/2}} \right) \left(3 \sum_{i=1}^{d} \partial^3_{iii} \log \tilde{q}_{t-1}(m_t) \Delta_t^i + \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \partial^3_{iij} \log \tilde{q}_{t-1}(m_t) \Delta_t^j \right) \\
&+ O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^3 \right), \ \forall \ell \geq 1.
\end{aligned}
$$

Here the last line follows because $(1-\alpha_t)^3 \left| \partial^3_{ijk} \log \tilde{q}_{t-1}(m_t) \right| = O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^3 \right)$ under Assumption 3 and $(1-\alpha_t)^3 \left\| \Delta_t \right\| = O_{\mathcal{L}^\ell(\tilde{Q}_t)}((1-\alpha_t)^3)$ under Assumption 4, both for all $\ell \geq 1$. The proof is now complete.

H.4 Proof of Lemma 4

Fix $x_t \in \mathbb{R}^d$. For brevity write $m_t = m_t(x_t)$, $\mu_t = \mu_t(x_t)$, and $\Delta_t = \Delta_t(x_t)$. Recall that

$$
T_p(\log \tilde{q}_{t-1}, x_{t-1}, m_t) = \frac{1}{p!} \sum_{\gamma \in \mathbb{N}^d : \sum_i \gamma^i = p} \partial^p_{\boldsymbol{a}} \log \tilde{q}_{t-1}(m_t) \prod_{i=1}^{d} (x^i_{t-1} - m^i_t)^{\gamma^i}
$$

where $\boldsymbol{a} \in [d]^p$ are the indices of differentiation in which the multiplicity of i is γ^i. First, for the expectation under $\tilde{P}_{t-1|t}$ (i.e., Gaussian centralized moments),

$$
\begin{aligned}
\mathbb{E}_{X_{t-1} \sim \tilde{P}_{t-1|t}} \left[\prod_{i=1}^{d} (X^i_{t-1} - m^i_t)^{\gamma^i} \right] &= \prod_{i=1}^{d} \mathbb{E}_{X_{t-1} \sim \tilde{P}_{t-1|t}} \left[(X^i_{t-1} - m^i_t)^{\gamma^i} \right] \\
&= \prod_{i=1}^{d} \left(\frac{1-\alpha_t}{\alpha_t} \right)^{\gamma^i/2} (\gamma^i - 1)!! \mathbb{1}\{\gamma^i \text{ is even}\} \\
&= \left(\frac{1-\alpha_t}{\alpha_t} \right)^{p/2} \prod_{i=1}^{d} (\gamma^i - 1)!! \mathbb{1}\{\gamma^i \text{ is even}\},
\end{aligned}
$$

where we use the convention that $(-1)!! = 1$. Next, for the expectation under $P_{t-1|t}$ (i.e., Gaussian non-centralized moments),

$$
\begin{aligned}
&\mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \left[\prod_{i=1}^{d} (X^i_{t-1} - m^i_t)^{\gamma^i} \right] \\
&= \prod_{i=1}^{d} \mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \left[(X^i_{t-1} - \mu^i_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}} \Delta^i_t)^{\gamma^i} \right] \\
&= \prod_{i=1}^{d} \sum_{\substack{\ell=0 \\ \ell \text{ even}}}^{\gamma^i} \binom{\gamma^i}{\ell} \left(-\frac{1-\alpha_t}{\sqrt{\alpha_t}} \Delta^i_t \right)^{\gamma^i - \ell} \left(\frac{1-\alpha_t}{\alpha_t} \right)^{\ell/2} (\ell - 1)!!
\end{aligned}
$$

To investigate their difference, we divide into the following few cases. Note that under Assumption 4, $(1-\alpha_t)^m \left\| \Delta_t(x_t) \right\| = O_{\mathcal{L}^\ell(\tilde{Q}_t)}((1-\alpha_t)^m)$ for any $m \geq 1/2$ and $\ell \geq 1$.

1. Case 1: p is even and all elements of γ^i are even. Then,

$$
\mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \left[\prod_{i=1}^{d} (X^i_{t-1} - m^i_t)^{\gamma^i} \right]
$$

$$= \prod_{i=1}^{d} \left(\left(\frac{1-\alpha_t}{\alpha_t} \right)^{\gamma^i/2} (\gamma^i - 1)!! + O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{\gamma^i/2+1} \right) \right)$$
$$= \left(\frac{1-\alpha_t}{\alpha_t} \right)^{p/2} \prod_{i=1}^{d} (\gamma^i - 1)!! + O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{p/2+1} \right)$$

2. Case 2: p is even and $\exists i^*$ such that γ^{i^*} is odd. Since $\sum_i \gamma^i = p$, there exists j^* such that γ^{j^*} is also odd. Then,

$$\mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \left[(X_{t-1}^{i^*} - m_t^{i^*})^{\gamma^{i^*}} \right] = O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{(\gamma^{i^*}+1)/2} \right),$$
$$\mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \left[(X_{t-1}^{j^*} - m_t^{j^*})^{\gamma^{j^*}} \right] = O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{(\gamma^{j^*}+1)/2} \right),$$

which implies that

$$\mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \left[\prod_{i=1}^{d} (X_{t-1}^i - m_t^i)^{\gamma^i} \right] = O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{p/2+1} \right)$$

3. Case 3: p is odd and $\exists i^*$ such that γ^{i^*} is odd. Then,

$$\mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \left[(X_{t-1}^{i^*} - m_t^{i^*})^{\gamma^{i^*}} \right] = O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{(\gamma^{i^*}+1)/2} \right),$$

which implies that

$$\mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \left[\prod_{i=1}^{d} (X_{t-1}^i - m_t^i)^{\gamma^i} \right] = O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{(p+1)/2} \right)$$

Combining these cases, we get

$$\left(\mathbb{E}_{X_{t-1} \sim \tilde{P}_{t-1|t}} - \mathbb{E}_{X_{t-1} \sim P_{t-1|t}} \right) \left[\prod_{i=1}^{d} (X_{t-1}^i - m_t^i)^{\gamma^i} \right]$$
$$= \begin{cases} O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{\frac{p}{2}+1} \right), & \forall p \geq 4 \text{ even} \\ O_{\mathcal{L}^\ell(\tilde{Q}_t)} \left((1-\alpha_t)^{\frac{p+1}{2}} \right), & \forall p \geq 4 \text{ odd} \end{cases}$$

The proof is complete by noting that the rate does not change when we take the expectation over $\tilde{Q}_t$ under Assumptions 3 and 4.

H.5 Proof of Lemma 5

Note that $\tilde{q}_{t|0}(x|x_0) = q_{t|0}(x|x_0)$ is the p.d.f. of $\mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I_d)$. Thus, the gradient of $\log \tilde{q}_t(x)$ equals

$$\nabla \log \tilde{q}_t(x) = \frac{\int_{x_0 \in \mathbb{R}^d} \nabla \tilde{q}_{t|0}(x|x_0) \mathrm{d}\tilde{Q}_0(x_0)}{\tilde{q}_t(x)} = -\frac{1}{1-\bar{\alpha}_t} \int_{x_0 \in \mathbb{R}^d} (x - \sqrt{\bar{\alpha}_t}x_0) \mathrm{d}\tilde{Q}_{0|t}(x_0|x). \quad (20)$$

Thus,

$$\nabla \log \tilde{q}_{t-1}(m_t) - \sqrt{\alpha_t} \nabla \log \tilde{q}_t(x_t)$$
$$= -\frac{1}{1-\bar{\alpha}_{t-1}} \int_{x_0 \in \mathbb{R}^d} (m_t - \sqrt{\bar{\alpha}_{t-1}}x_0) \mathrm{d}\tilde{Q}_{0|t-1}(x_0|m_t) + \frac{\sqrt{\alpha_t}}{1-\bar{\alpha}_t} \int_{x_0 \in \mathbb{R}^d} (x_t - \sqrt{\bar{\alpha}_t}x_0) \mathrm{d}\tilde{Q}_{0|t}(x_0|x_t)$$
$$= -\frac{1}{1-\bar{\alpha}_t} \Bigg(\left(\frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}} - 1 \right) \int_{x_0 \in \mathbb{R}^d} (m_t - \sqrt{\bar{\alpha}_{t-1}}x_0) \mathrm{d}\tilde{Q}_{0|t-1}(x_0|m_t)$$
$$+ \int_{x_0 \in \mathbb{R}^d} (m_t - \sqrt{\bar{\alpha}_{t-1}}x_0) \mathrm{d}\tilde{Q}_{0|t-1}(x_0|m_t) - \sqrt{\alpha_t} \int_{x_0 \in \mathbb{R}^d} (x_t - \sqrt{\bar{\alpha}_t}x_0) \mathrm{d}\tilde{Q}_{0|t}(x_0|x_t) \Bigg)$$

$$
\begin{aligned}
&\overset{(i)}{=} \frac{1}{1-\bar{\alpha}_t}\Big((1-\bar{\alpha}_t-(1-\bar{\alpha}_{t-1}))\nabla \log \tilde{q}_{t-1}(m_t)\Big) \\
&\quad - \frac{1}{1-\bar{\alpha}_t}\left(\int_{x_0\in\mathbb{R}^d} (m_t-\sqrt{\bar{\alpha}_{t-1}}x_0)\mathrm{d}\tilde{Q}_{0|t-1}(x_0|m_t) - \sqrt{\alpha_t}\int_{x_0\in\mathbb{R}^d}(x_t-\sqrt{\bar{\alpha}_t}x_0)\mathrm{d}\tilde{Q}_{0|t}(x_0|x_t)\right) \\
&= \underbrace{\frac{\bar{\alpha}_{t-1}(1-\alpha_t)}{1-\bar{\alpha}_t}\nabla\log\tilde{q}_{t-1}(m_t)}_{\text{term 1}} - \underbrace{\frac{1}{1-\bar{\alpha}_t}(m_t-\sqrt{\alpha_t}x_t)}_{\text{term 2}} + \underbrace{\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\int_{x_0\in\mathbb{R}^d} x_0\mathrm{d}\tilde{Q}_{0|t}(x_0|x_t)}_{\text{term 3}} \\
&\quad + \underbrace{\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\left(\int_{x_0\in\mathbb{R}^d} x_0\mathrm{d}\tilde{Q}_{0|t-1}(x_0|m_t) - \int_{x_0\in\mathbb{R}^d} x_0\mathrm{d}\tilde{Q}_{0|t}(x_0|x_t)\right)}_{\text{term 4}} \qquad (21)
\end{aligned}
$$

where (i) follows from Tweedie's formula. Among the four terms in (21), the first term satisfies that

$$
\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\frac{\bar{\alpha}_{t-1}(1-\alpha_t)}{1-\bar{\alpha}_t}\nabla\log\tilde{q}_{t-1}(m_t)\right\|^2 \lesssim \frac{d(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2(1-\bar{\alpha}_{t-1})}
$$

by (Liang et al., 2024, Lemma 17). In the second term in (21), by Tweedie's formula,

$$
\begin{aligned}
m_t-\sqrt{\alpha_t}x_t &= \frac{x_t}{\sqrt{\alpha_t}} + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log\tilde{q}_t(x_t) - \sqrt{\alpha_t}x_t \\
&= \frac{1-\alpha_t}{\sqrt{\alpha_t}}(x_t+\nabla\log\tilde{q}_t(x_t)).
\end{aligned}
$$

Thus, by (Liang et al., 2024, Lemma 15) and Assumption 1, the second term satisfies that

$$
\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\frac{1}{1-\bar{\alpha}_t}(m_t-\sqrt{\alpha_t}x_t)\right\|^2 \lesssim \frac{d(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^3}.
$$

The third term in (21) satisfies that

$$
\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\int_{x_0\in\mathbb{R}^d} x_0\mathrm{d}\tilde{Q}_{0|t}(x_0|x_t)\right\|^2 \lesssim \frac{d(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2}
$$

by Jensen's inequality and Assumption 1.

To deal with the last term in (21), note that

$$
\begin{aligned}
\mathrm{d}\tilde{Q}_{0|t-1}(x_0|m_t) &= \frac{\tilde{q}_{t-1|0}(m_t|x_0)}{\tilde{q}_{t-1}(m_t)}\mathrm{d}\tilde{Q}_0(x_0) = \frac{\tilde{q}_{t-1|0}(m_t|x_0)}{\int_{y\in\mathbb{R}^d}\tilde{q}_{t-1|0}(m_t|y)\mathrm{d}\tilde{Q}_0(y)}\mathrm{d}\tilde{Q}_0(x_0), \\
\mathrm{d}\tilde{Q}_{0|t}(x_0|x_t) &= \frac{\tilde{q}_{t|0}(x_t|x_0)}{\tilde{q}_t(x_t)}\mathrm{d}\tilde{Q}_0(x_0) = \frac{\tilde{q}_{t|0}(x_t|x_0)}{\int_{y\in\mathbb{R}^d}\tilde{q}_{t|0}(x_t|y)\mathrm{d}\tilde{Q}_0(y)}\mathrm{d}\tilde{Q}_0(x_0).
\end{aligned}
$$

Thus, the last term in (21) is equal to

$$
\begin{aligned}
&\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\left(\int_{x_0\in\mathbb{R}^d} x_0\mathrm{d}\tilde{Q}_{0|t-1}(x_0|m_t) - \int_{x_0\in\mathbb{R}^d} x_0\mathrm{d}\tilde{Q}_{0|t}(x_0|x_t)\right) \\
&= \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\cdot\frac{1}{\tilde{q}_{t-1}(m_t)\tilde{q}_t(x_t)}\left(\int_{x,y\in\mathbb{R}^d} x(\tilde{q}_{t-1|0}(m_t|x)\tilde{q}_{t|0}(x_t|y) - \tilde{q}_{t|0}(x_t|x)\tilde{q}_{t-1|0}(m_t|y))\mathrm{d}\tilde{Q}_0(x)\mathrm{d}\tilde{Q}_0(y)\right)
\end{aligned}
$$

where

$$
\begin{aligned}
&\tilde{q}_{t-1|0}(m_t|x)\tilde{q}_{t|0}(x_t|y) - \tilde{q}_{t|0}(x_t|x)\tilde{q}_{t-1|0}(m_t|y) \\
&= \tilde{q}_{t|0}(x_t|x)\tilde{q}_{t-1|0}(m_t|y)\left(\frac{\tilde{q}_{t-1|0}(m_t|x)\tilde{q}_{t|0}(x_t|y)}{\tilde{q}_{t|0}(x_t|x)\tilde{q}_{t-1|0}(m_t|y)} - 1\right) \\
&= \tilde{q}_{t|0}(x_t|x)\tilde{q}_{t-1|0}(m_t|y)\times \\
&\quad\left(\exp\left(-\frac{\|m_t-\sqrt{\bar{\alpha}_{t-1}}x\|^2}{2(1-\bar{\alpha}_{t-1})} - \frac{\|x_t-\sqrt{\bar{\alpha}_t}y\|^2}{2(1-\bar{\alpha}_t)} + \frac{\|m_t-\sqrt{\bar{\alpha}_{t-1}}y\|^2}{2(1-\bar{\alpha}_{t-1})} + \frac{\|x_t-\sqrt{\bar{\alpha}_t}x\|^2}{2(1-\bar{\alpha}_t)}\right) - 1\right)
\end{aligned}
$$

$=: \tilde{q}_{t|0}(x_t|x)\tilde{q}_{t-1|0}(m_t|y)(e^{\Delta}-1)$

in which we have defined the exponent as Δ. Now,

$$\begin{aligned}
\Delta &= -\frac{\|m_t - \sqrt{\bar{\alpha}_{t-1}}x\|^2}{2(1-\bar{\alpha}_{t-1})} - \frac{\|x_t - \sqrt{\bar{\alpha}_t}y\|^2}{2(1-\bar{\alpha}_t)} + \frac{\|m_t - \sqrt{\bar{\alpha}_{t-1}}y\|^2}{2(1-\bar{\alpha}_{t-1})} + \frac{\|x_t - \sqrt{\bar{\alpha}_t}x\|^2}{2(1-\bar{\alpha}_t)} \\
&= \frac{\sqrt{\bar{\alpha}_{t-1}}(x-y)^\intercal m_t + \bar{\alpha}_{t-1}\|y\|^2 - \bar{\alpha}_{t-1}\|x\|^2}{2(1-\bar{\alpha}_{t-1})} - \frac{\sqrt{\bar{\alpha}_t}(x-y)^\intercal x_t + \bar{\alpha}_t\|y\|^2 - \bar{\alpha}_t\|x\|^2}{2(1-\bar{\alpha}_t)} \\
&= \frac{1}{2}\left(\frac{\sqrt{\bar{\alpha}_t}(1-\alpha_t)}{\alpha_t(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}\nabla\log\tilde{q}_t(x_t)\right)^\intercal (x-y) \\
&\quad + \frac{\bar{\alpha}_{t-1}(1-\alpha_t)}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)}(\|y\|^2 - \|x\|^2).
\end{aligned}$$

Now, with the α_t defined in (8), following from Lemma 6,

$$\begin{aligned}
\frac{\sqrt{\bar{\alpha}_t}(1-\alpha_t)}{\alpha_t(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)} = O\left(\frac{1-\alpha_t}{(1-\bar{\alpha}_{t-1})^2}\right) = O\left(\frac{\log T}{T}\right), \\
\frac{1-\alpha_t}{1-\bar{\alpha}_{t-1}} = O\left(\frac{\log T}{T}\right), \\
\frac{\bar{\alpha}_{t-1}(1-\alpha_t)}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)} = O\left(\frac{1-\alpha_t}{(1-\bar{\alpha}_{t-1})^2}\right) = O\left(\frac{\log T}{T}\right).
\end{aligned}$$

Thus, for fixed x, y, x_t, $\Delta \to 0$ as $T \to \infty$, and thus when T becomes large,

$$e^{\Delta} - 1 = \Delta + O(\Delta^2) \lesssim |\Delta|, \quad \forall x_t \in \mathbb{R}^d.$$

Also, since $\tilde{q}_{t|0}(x_t|x)$ and $\tilde{q}_{t-1|0}(m_t|y)$ decay exponentially in terms of x and y (for any fixed x_t), we have

$$\begin{aligned}
\int \tilde{q}_{t|0}(x_t|x)\mathrm{poly}(x)\mathrm{d}\tilde{Q}_0(x) < \infty, \\
\int \tilde{q}_{t-1|0}(m_t|y)\mathrm{poly}(y)\mathrm{d}\tilde{Q}_0(y) < \infty.
\end{aligned}$$

Thus, the limit and the integral can be exchanged due to Dominated Convergence Theorem. Thus, the fourth term in (21) gives us

$$\begin{aligned}
&\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\left(\int_{x_0\in\mathbb{R}^d} x_0 \mathrm{d}\tilde{Q}_{0|t-1}(x_0|m_t) - \int_{x_0\in\mathbb{R}^d} x_0\mathrm{d}\tilde{Q}_{0|t}(x_0|x_t)\right) \\
&\lesssim \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\cdot\frac{1}{\tilde{q}_{t-1}(m_t)\tilde{q}_t(x_t)}\left(\int_{x,y\in\mathbb{R}^d} x\tilde{q}_{t|0}(x_t|x)\tilde{q}_{t-1|0}(m_t|y)\,|\Delta|\,\mathrm{d}\tilde{Q}_0(x)\mathrm{d}\tilde{Q}_0(y)\right) \\
&= \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\left(\int_{x,y\in\mathbb{R}^d}(x\cdot|\Delta|)\mathrm{d}\tilde{Q}_{0|t}(x|x_t)\mathrm{d}\tilde{Q}_{0|t-1}(y|m_t)\right)
\end{aligned}$$

and, from definition of Δ and using Cauchy-Schwartz and Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\int_{x,y\in\mathbb{R}^d}(x\cdot|\Delta|)\mathrm{d}\tilde{Q}_{0|t}(x|X_t)\mathrm{d}\tilde{Q}_{0|t-1}(y|m_t(X_t))\right\|^2 \\
&\lesssim \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_{t-1})^2(1-\bar{\alpha}_t)^4}\cdot \\
&\quad \mathbb{E}_{\substack{X_t\sim\tilde{Q}_t \\ X\sim\tilde{Q}_{0|t}(\cdot|X_t) \\ Y\sim\tilde{Q}_{0|t-1}(\cdot|m_t(X_t))}}\Bigg[\left\|\sqrt{\bar{\alpha}_t}X\right\|^2\Bigg((\|X_t\|^2 + (1-\bar{\alpha}_t)^2\|\nabla\log\tilde{q}_t(X_t)\|^2) \\
&\quad (\left\|\sqrt{\bar{\alpha}_t}X\right\|^2 + \|\sqrt{\bar{\alpha}_{t-1}}Y\|^2) + (\left\|\sqrt{\bar{\alpha}_t}X\right\|^4 + \|\sqrt{\bar{\alpha}_{t-1}}Y\|^4)\Bigg)\Bigg] \\
&= \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_{t-1})^2(1-\bar{\alpha}_t)^4}\cdot
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}_{\substack{X_t\sim\tilde{Q}_t\\ X\sim\tilde{Q}_{0|t}(\cdot|X_t)\\ Y\sim\tilde{Q}_{0|t-1}(\cdot|m_t(X_t))}}\Big[\left\|\sqrt{\bar{\alpha}_t}X\right\|^4\left(\|X_t\|^2+(1-\bar{\alpha}_t)^2\left\|\nabla\log\tilde{q}_t(X_t)\right\|^2\right)\\
&+\left\|\sqrt{\bar{\alpha}_t}X\right\|^2\left\|\sqrt{\bar{\alpha}_{t-1}}Y\right\|^2\left(\|X_t\|^2+(1-\bar{\alpha}_t)^2\left\|\nabla\log\tilde{q}_t(X_t)\right\|^2\right)\\
&+\left\|\sqrt{\bar{\alpha}_t}X\right\|^6+\left\|\sqrt{\bar{\alpha}_t}X\right\|^2\left\|\sqrt{\bar{\alpha}_{t-1}}Y\right\|^4\Big]\\
\le\ &\frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_{t-1})^2(1-\bar{\alpha}_t)^4}\cdot\\
&\left(\mathbb{E}_{X\sim\tilde{Q}_0}\left\|\sqrt{\bar{\alpha}_t}X\right\|^6\right)^{2/3}\left(\mathbb{E}_{X_t\sim\tilde{Q}_t}(\|X_t\|^6+(1-\bar{\alpha}_t)^6\left\|\nabla\log\tilde{q}_t(X_t)\right\|^6)\right)^{1/3}\\
&+\left(\mathbb{E}_{X\sim\tilde{Q}_0}\left\|\sqrt{\bar{\alpha}_t}X\right\|^6\right)^{1/3}\left(\mathbb{E}_{\substack{X_t\sim\tilde{Q}_t\\ Y\sim\tilde{Q}_{0|t-1}(\cdot|m_t(X_t))}}\left\|\sqrt{\bar{\alpha}_{t-1}}Y\right\|^6\right)^{1/3}\\
&\quad\left(\mathbb{E}_{X_t\sim\tilde{Q}_t}(\|X_t\|^6+(1-\bar{\alpha}_t)^6\left\|\nabla\log\tilde{q}_t(X_t)\right\|^6)\right)^{1/3}\\
&+\mathbb{E}_{X\sim\tilde{Q}_0}\left\|\sqrt{\bar{\alpha}_t}X\right\|^6+\left(\mathbb{E}_{X\sim\tilde{Q}_0}\left\|\sqrt{\bar{\alpha}_t}X\right\|^6\right)^{1/3}\left(\mathbb{E}_{\substack{X_t\sim\tilde{Q}_t\\ Y\sim\tilde{Q}_{0|t-1}(\cdot|m_t(X_t))}}\left\|\sqrt{\bar{\alpha}_{t-1}}Y\right\|^6\right)^{2/3}\\
\overset{(ii)}{\lesssim}\ &\frac{d^3(1-\alpha_t)^2}{(1-\bar{\alpha}_{t-1})^2(1-\bar{\alpha}_t)^4},
\end{aligned}
$$

where (ii) follows because, following (Liang et al., 2024, Lemmas 15–17) and by the lemma assumption that $\mathbb{E}_{X_0\sim\tilde{Q}_0}\|X_0\|^6\lesssim d^3$, we have

$$
\begin{aligned}
&\mathbb{E}_{X\sim\tilde{Q}_0}\left\|\sqrt{\bar{\alpha}_t}X\right\|^6\lesssim d^3,\\
&\mathbb{E}_{X_t\sim\tilde{Q}_t}\|X_t\|^6\le\mathbb{E}_{X_0\sim\tilde{Q}_0}\left\|\sqrt{\bar{\alpha}_t}X_0\right\|^6+(1-\bar{\alpha}_t)^3\mathbb{E}_{\bar{W}\sim\mathcal{N}(0,I_d)}\left\|\bar{W}\right\|^6\lesssim d^3,\\
&\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\nabla\log\tilde{q}_t(X_t)\right\|^6\lesssim\frac{d^3}{(1-\bar{\alpha}_t)^3},\\
&\mathbb{E}_{\substack{X_t\sim\tilde{Q}_t\\ Y\sim\tilde{Q}_{0|t-1}(\cdot|m_t(X_t))}}\left\|\sqrt{\bar{\alpha}_{t-1}}Y\right\|^6\\
&\quad\le\mathbb{E}_{\substack{X_t\sim\tilde{Q}_t\\ Y\sim\tilde{Q}_{0|t-1}(\cdot|m_t(X_t))}}\left\|m_t-\sqrt{\bar{\alpha}_{t-1}}Y\right\|^6+\mathbb{E}_{X_t\sim\tilde{Q}_t}\|m_t\|^6\lesssim d^3.
\end{aligned}
$$

Hence, combining the rates of all parts, we obtain that

$$
(1-\alpha_t)\sqrt{\mathbb{E}_{X_t\sim\tilde{Q}_t}\left\|\nabla\log\tilde{q}_{t-1}(m_t(X_t))-\sqrt{\alpha_t}\nabla\log\tilde{q}_t(X_t)\right\|^2}\lesssim\frac{d^{3/2}(1-\alpha_t)^2}{(1-\bar{\alpha}_{t-1})^3}.
$$

H.6 Lemma 6 and its proof

Lemma 6. *The α_t defined in* (8) *(with $c>1$) satisfy*

$$
\frac{1-\alpha_t}{(1-\bar{\alpha}_{t-1})^p}\lesssim\frac{\log T\log(1/\delta)}{\delta^{p-1}T}\ \text{ while }\ \bar{\alpha}_T=o(T^{-1}),\quad\forall 2\le t\le T,\ p\ge 1.
$$

Proof. The proof is similar to that of (Li et al., 2024c, Eq (39)). We first prove the second relationship. First, note that if T is large,

$$
\delta\left(1+\frac{c\log T}{T}\right)^{\frac{T}{\log T}}\asymp\delta e^c>1.
$$

Thus, with any fixed $r\in(0,1)$ such that $t\ge rT\ (\ge\frac{T}{\log T})$, we have

$$
1-\alpha_t=\frac{c\log T}{T}\min\left\{\delta\left(1+\frac{c\log T}{T}\right)^t,1\right\}=\frac{c\log T}{T}.
$$

As a result,

$$\bar{\alpha}_T \le \prod_{t=\lfloor rT \rfloor}^{T} \alpha_t = \left(1 - \frac{c\log T}{T}\right)^{\lceil (1-r)T \rceil} \asymp \exp\left(\lceil (1-r)T \rceil \left(-\frac{c\log T}{T}\right)\right) = O(T^{-(1-r)c}). \tag{22}$$

Given any $c > 1$, we can always find some r such that $(1-r)c > 1$ (say, $r = (c-1)/2$ if $c \in (1,2)$ and $r = 1/4$ if $c \ge 2$). This shows that α_t satisfies $\bar{\alpha}_T = o\left(T^{-1}\right)$ if $c > 1$.

Now, for the first relatinoship, define τ such that

$$\delta\left(1 + \frac{c\log T}{T}\right)^{\tau} \le 1 < \delta\left(1 + \frac{c\log T}{T}\right)^{\tau+1}. \tag{23}$$

Here τ is unique since $1 - \alpha_t$ is non-decreasing. In other words, τ is the last time that $1 - \alpha_t$ is exponentially growing. Assume that T is large enough such that $\tau \ge 2$. Below, we show that

$$1 - \bar{\alpha}_{t-1} \ge \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t}, \quad \forall 2 \le t \le \tau. \tag{24}$$

If $t = 2$,

$$1 - \bar{\alpha}_{t-1} = 1 - \bar{\alpha}_1 = 1 - \alpha_1 = \delta \ge \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right).$$

Here the last inequality holds when T is sufficiently large. For $t > 2$, suppose for purpose of contradiction that there exists $2 < t_0 \le \tau$ such that

$$1 - \bar{\alpha}_{t_0-1} < \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t_0} \text{ while } 1 - \bar{\alpha}_{t-1} \ge \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t}, \ \forall 2 \le t \le t_0 - 1.$$

In words, t_0 is defined as the ***first*** time that (24) is violated. To arrive at a contradiction, we first write

$$\begin{aligned} 1 - \bar{\alpha}_{t_0-1} &= (1 - \bar{\alpha}_{t_0-2})\left(1 + \frac{\bar{\alpha}_{t_0-2}(1-\alpha_{t_0-1})}{1 - \bar{\alpha}_{t_0-2}}\right) \\ &\ge \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t_0-1}\left(1 + \frac{\bar{\alpha}_{t_0-2}(1-\alpha_{t_0-1})}{1 - \bar{\alpha}_{t_0-2}}\right). \end{aligned}$$

Here the inequality holds because t_0 is the first time that (24) is violated, and thus (24) stills holds for $t = t_0 - 1$. Also,

$$1 - \bar{\alpha}_{t_0-2} \le 1 - \bar{\alpha}_{t_0-1} \overset{(i)}{<} \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t_0} \overset{(ii)}{\le} \frac{1}{2}\delta\left(1 + \frac{c\log T}{T}\right)^{t_0-1} \overset{(iii)}{\le} \frac{1}{2}$$

where (i) holds because (24) is violated at $t = t_0$, (ii) holds when T is sufficiently large, and (iii) holds because $t_0 - 1 \le \tau$ and by the definition of τ in (23). Thus,

$$\frac{\bar{\alpha}_{t_0-2}(1-\alpha_{t_0-1})}{1 - \bar{\alpha}_{t_0-2}} \ge \frac{\frac{1}{2}\frac{c\log T}{T}\delta\left(1 + \frac{c\log T}{T}\right)^{t_0-1}}{\frac{1}{2}\delta\left(1 + \frac{c\log T}{T}\right)^{t_0-1}} = \frac{c\log T}{T},$$

and thus

$$1 - \bar{\alpha}_{t_0-1} \ge \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t_0-1}\left(1 + \frac{\bar{\alpha}_{t_0-2}(1-\alpha_{t_0-1})}{1 - \bar{\alpha}_{t_0-2}}\right) \ge \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t_0}.$$

We have reached a contradiction. Therefore, we have shown that (24) holds.

Now, (24) implies that

$$1 - \bar{\alpha}_{t-1} \ge \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t} \ge \frac{1}{3}\delta\left(1 + \frac{c\log T}{T}\right)^{t/p}, \quad \forall 2 \le t \le \tau.$$

There are two cases:

- If $2 \le t \le \tau$, then

$$\frac{1-\alpha_t}{(1-\bar{\alpha}_{t-1})^p} \le \frac{\frac{c\log T}{T}\delta\left(1+\frac{c\log T}{T}\right)^t}{\frac{1}{3^p}\delta^p\left(1+\frac{c\log T}{T}\right)^t} = \frac{3^p c\log T}{\delta^{p-1}T}.$$

- If $t > \tau$, then

$$\frac{1-\alpha_t}{(1-\bar{\alpha}_{t-1})^p} \le \frac{1-\alpha_t}{(1-\bar{\alpha}_{\tau-1})^p} \le \frac{\frac{c\log T}{T}}{\frac{1}{3^p}\delta^p\left(1+\frac{c\log T}{T}\right)^\tau} = \frac{\frac{c\log T}{T}\left(1+\frac{c\log T}{T}\right)}{3^{-p}\delta^{p-1}\left(1+\frac{c\log T}{T}\right)^{\tau+1}}$$
$$< \frac{3^p c\log T}{\delta^{p-1}T}\left(1+\frac{c\log T}{T}\right).$$

In both cases, if T is large enough, noting that $c \gtrsim \log(1/\delta)$, we have

$$\frac{1-\alpha_t}{(1-\bar{\alpha}_{t-1})^p} \le \frac{4^p c\log T}{\delta^{p-1}T} \lesssim \frac{\log T\log(1/\delta)}{\delta^{p-1}T}, \quad \forall 2 \le t \le T$$

because p and c are constants (that do not depend on T, d, and δ). The proof is now complete. □

H.7 Lemma 7 and its proof

Lemma 7. *With the α_t defined in* (8)*, given any $p > 0$, if $\delta p < 1$,*

$$\sum_{t=2}^{T}(1-\alpha_t)\bar{\alpha}_t^p \le \left(\frac{1}{p}(1-\delta)^p e^{-p\delta\log(1/\delta)} + (1-\delta)^p\frac{e^{-p\delta\log(1/\delta)}-1}{1-\delta p}\right)\left(1+O\left(\frac{\log T}{T}\right)\right).$$

Further, when $\delta \ll 1$,

$$\sum_{t=2}^{T}(1-\alpha_t)\bar{\alpha}_t^p \le \frac{1}{p} - \left(1+\frac{p+1}{2p}\right)\frac{c\log T}{T} + \tilde{O}\left(\frac{1}{T^2}\right).$$

Proof. Define the sum as s_T. Recall that

$$1-\alpha_1 = \delta, \quad 1-\alpha_t = \frac{c\log T}{T}\min\left\{\delta\left(1+\frac{c\log T}{T}\right)^t, 1\right\}, \ \forall 2 \le t \le T.$$

We first note a relationship that for fixed $\delta \ne 0$ and $p > 0$. As $z \to \infty$,

$$(1-\delta z^{-1})^{pz} = e^{pz\log(1-\delta z^{-1})} = e^{pz(-\delta z^{-1}+\delta^2 z^{-2}/2+O(z^{-3}))} = e^{-\delta p}(1+\delta^2 p z^{-1}/2) + O(z^{-2}). \tag{25}$$

We also use the fact from binomial series that

$$(1-z^{-1})^p = 1 - pz^{-1} + \frac{p(p-1)}{2}z^{-2} + O(z^{-3}). \tag{26}$$

Define $t^* := \sup\left\{t \in [1,T] : \delta\left(1+\frac{c\log T}{T}\right)^t \le 1\right\}$. Thus, $\alpha_t \equiv 1 - \frac{c\log T}{T}$ for all $t > t^*$. Note that when T becomes large, $t^* = \Theta\left(\frac{T}{\log T}\right)$. To further understand the big-Θ term, note that using (25),

$$\delta\left(1+\frac{c\log T}{T}\right)^{\frac{T\log(1/\delta)}{c\log T}-\log(1/\delta)}$$
$$= \left(1+\frac{\log(1/\delta)c\log T}{2T}+\tilde{O}\left(\frac{1}{T^2}\right)\right)\left(1-\frac{\log(1/\delta)c\log T}{T}+\tilde{O}\left(\frac{1}{T^2}\right)\right)$$
$$= 1-\frac{\log(1/\delta)c\log T}{2T}+\tilde{O}\left(\frac{1}{T^2}\right)$$

< 1 as $T \to \infty$.

This implies that

$$t^* \geq \log(1/\delta) \left(\frac{T}{c \log T} - 1 \right). \tag{27}$$

To start, we suppose $T > t^*$ is large enough and decompose the sum as

$$\begin{aligned} s_T &= \sum_{t=2}^{t^*} (1-\alpha_t)\bar{\alpha}_t^p + \sum_{t=t^*+1}^{T} (1-\alpha_t)\bar{\alpha}_t^p \\ &= \frac{c\log T}{T}\delta(1-\delta)^p \sum_{t=2}^{t^*} \left(1 + \frac{c\log T}{T}\right)^t \prod_{i=2}^{t} \left(1 - \delta\frac{c\log T}{T}\left(1+\frac{c\log T}{T}\right)^i\right)^p \\ &\quad + \bar{\alpha}_{t^*}^p \frac{c\log T}{T} \sum_{t=t^*+1}^{T} \left(1 - \frac{c\log T}{T}\right)^{p(t-t^*)}. \end{aligned} \tag{28}$$

Now we first focus on the second term in (28).

$$\begin{aligned} &\frac{c\log T}{T} \sum_{t=t^*+1}^{T} \left(1 - \frac{c\log T}{T}\right)^{p(t-t^*)} \\ &= \left(1 - \frac{c\log T}{T}\right)^p \frac{c\log T}{T} \cdot \frac{1 - \left(1 - \frac{c\log T}{T}\right)^{p(T-t^*)}}{1 - \left(1 - \frac{c\log T}{T}\right)^p} \\ &\overset{(i)}{=} \left(1 - \frac{c\log T}{T}\right)^p \frac{c\log T}{T} \cdot \frac{1 - \left(1 - \frac{c\log T}{T}\right)^{p(T-t^*)}}{1 - \left(1 - \frac{pc\log T}{T} + p(p-1)\frac{c^2(\log T)^2}{2T^2} + \tilde{O}\left(\frac{1}{T^3}\right)\right)} \\ &\overset{(ii)}{=} \frac{1}{p}\left(1 - \frac{pc\log T}{T} + \tilde{O}\left(\frac{1}{T^2}\right)\right)\left(1 + \frac{(p-1)c\log T}{2T} + \tilde{O}\left(\frac{1}{T^2}\right) - O\left(\frac{1}{T^{pc/2}}\right)\right) \\ &= \frac{1}{p}\left(1 - \frac{(p+1)c\log T}{2T}\right) + \tilde{O}\left(\frac{1}{T^2}\right) \end{aligned}$$

where (i) follows from (26), and (ii) is because $t^* = \Theta(T/\log T)$ and thus $T - t^* > T/2$ for large T. Also, for all $t = 2, \ldots, t^*$,

$$\begin{aligned} \bar{\alpha}_t^p &= (1-\delta)^p \prod_{i=2}^{t} \left(1 - \delta\frac{c\log T}{T}\left(1+\frac{c\log T}{T}\right)^i\right)^p \\ &\leq (1-\delta)^p \left(1 - \delta\frac{c\log T}{T}\right)^{p(t-1)} \end{aligned}$$

Thus, we have

$$\begin{aligned} \bar{\alpha}_{t^*}^p &\leq (1-\delta)^p \left(1 - \delta\frac{c\log T}{T}\right)^{p(t^*-1)} \\ &= (1-\delta)^p \left(1 + \delta p\frac{c\log T}{T} + \tilde{O}\left(\frac{1}{T^2}\right)\right)\left(1 - \delta\frac{c\log T}{T}\right)^{pt^*} \\ &\overset{(iii)}{\leq} (1-\delta)^p \left(1 + \delta p(1+\log(1/\delta))\frac{c\log T}{T}\right) \times \\ &\qquad e^{-p\delta\log(1/\delta)}\left(1 + \delta^2 p\log(1/\delta)\frac{c\log T}{2T}\right) + \tilde{O}\left(\frac{1}{T^2}\right) \\ &= (1-\delta)^p e^{-p\delta\log(1/\delta)} \left(1 + \delta p\left(1 + \log(1/\delta) + (\delta/2)\log(1/\delta)\right)\frac{c\log T}{T}\right) + \tilde{O}\left(\frac{1}{T^2}\right). \end{aligned}$$

Here (iii) follows because using (27) and (25), we have that

$$\left(1-\delta\frac{c\log T}{T}\right)^{pt^*} \le \left(1-\delta\frac{c\log T}{T}\right)^{p\log(1/\delta)\left(\frac{T}{c\log T}-1\right)}$$
$$= \left(1+\delta p\log(1/\delta)\frac{c\log T}{T}\right)e^{-p\delta\log(1/\delta)}\left(1+\delta^2 p\log(1/\delta)\frac{c\log T}{2T}\right)+\tilde{O}\left(\frac{1}{T^2}\right). \tag{29}$$

Thus, the second term in (28) satisfies that

$$\sum_{t=t^*+1}^{T}(1-\alpha_t)\bar{\alpha}_t^p$$
$$\le \frac{1}{p}(1-\delta)^p e^{-p\delta\log(1/\delta)}\left(1-\left(\frac{p+1}{2}-\delta p\left(1+\log(1/\delta)+(\delta/2)\log(1/\delta)\right)\right)\frac{c\log T}{T}\right)$$
$$+\tilde{O}\left(\frac{1}{T^2}\right). \tag{30}$$

Now we turn to the first term in (28), in which the summation can be upper-bounded as

$$\sum_{t=2}^{t^*}\left(1+\frac{c\log T}{T}\right)^t\prod_{i=2}^{t}\left(1-\delta\frac{c\log T}{T}\left(1+\frac{c\log T}{T}\right)^i\right)^p$$
$$\le \sum_{t=2}^{t^*}\left(1+\frac{c\log T}{T}\right)^t\left(1-\delta\frac{c\log T}{T}\right)^{p(t-1)} =: \left(1+\frac{c\log T}{T}\right)\sum_{t=1}^{t^*-1}q^t$$

where

$$q := \left(1+\frac{c\log T}{T}\right)\left(1-\delta\frac{c\log T}{T}\right)^p$$
$$= \left(1+\frac{c\log T}{T}\right)\left(1-\delta p\frac{c\log T}{T}+\delta^2 p(p-1)\frac{c^2(\log T)^2}{2T^2}+\tilde{O}\left(\frac{1}{T^3}\right)\right)$$
$$= 1+(1-\delta p)\frac{c\log T}{T}+\left(\frac{\delta^2 p(p-1)}{2}-\delta p\right)\frac{c^2(\log T)^2}{T^2}+\tilde{O}\left(\frac{1}{T^3}\right).$$

Note that by assumption $\delta p<1$. Also, by definition of t^* and (29),

$$\delta q^{t^*} \le \left(1-\delta\frac{c\log T}{T}\right)^{pt^*}$$
$$\le e^{-p\delta\log(1/\delta)}\left(1+\delta p\log(1/\delta)(1+\delta/2)\frac{c\log T}{T}\right)+\tilde{O}\left(\frac{1}{T^2}\right).$$

Thus, we have

$$\delta\frac{c\log T}{T}\sum_{t=1}^{t^*-1}q^t = \frac{c\log T}{T}\times\frac{\delta q^{t^*}-\delta q}{q-1}$$
$$\le \frac{c\log T}{T}\times\frac{e^{-p\delta\log(1/\delta)}\left(1+\delta p\log(1/\delta)(1+\delta/2)\frac{c\log T}{T}\right)+\tilde{O}\left(\frac{1}{T^2}\right)-\delta q}{q-1}$$
$$\overset{(iv)}{=} \frac{c\log T}{T}\times\frac{e^{-p\delta\log(1/\delta)}\left(1+\delta p\log(1/\delta)(1+\delta/2)\frac{c\log T}{T}\right)+\tilde{O}\left(\frac{1}{T^2}\right)-1-(1-\delta p)\frac{c\log T}{T}+\tilde{O}\left(\frac{1}{T^2}\right)}{(1-\delta p)\frac{c\log T}{T}+\left(\frac{\delta^2 p(p-1)}{2}-\delta p\right)\frac{c^2(\log T)^2}{T^2}+\tilde{O}\left(\frac{1}{T^3}\right)}$$
$$= \left(\frac{e^{-p\delta\log(1/\delta)}-1}{1-\delta p}+\left(\frac{\delta p\log(1/\delta)(1+\delta/2)}{1-\delta p}-1\right)\frac{c\log T}{T}\right)\times$$
$$\left(1+\frac{\delta p-\frac{\delta^2 p(p-1)}{2}}{1-\delta p}\cdot\frac{c\log T}{T}\right)+\tilde{O}\left(\frac{1}{T^2}\right)$$

$$= \frac{e^{-p\delta\log(1/\delta)}-1}{1-\delta p} + \left(\frac{\delta p\log(1/\delta)(1+\delta/2)}{1-\delta p} - 1 + \frac{e^{-p\delta\log(1/\delta)}-1}{1-\delta p}\cdot\frac{\delta p(1-\delta(p-1)/2)}{1-\delta p}\right)\frac{c\log T}{T} + \tilde{O}\left(\frac{1}{T^2}\right).$$

where (iv) follows from (26). Therefore,

$$\begin{aligned}
\sum_{t=2}^{t^*}(1-\alpha_t)\bar{\alpha}_t^p &= (1-\delta)^p\left(1+\frac{c\log T}{T}\right)\left(\delta\frac{c\log T}{T}\sum_{t=1}^{t^*-1}q^t\right)\\
&\le (1-\delta)^p\frac{e^{-p\delta\log(1/\delta)}-1}{1-\delta p}\\
&\quad + (1-\delta)^p\Bigg(\frac{\delta p\log(1/\delta)(1+\delta/2)}{1-\delta p} - 1\\
&\qquad + \frac{e^{-p\delta\log(1/\delta)}-1}{1-\delta p}\left(1+\frac{\delta p(1-\delta(p-1)/2)}{1-\delta p}\right)\Bigg)\frac{c\log T}{T} + \tilde{O}\left(\frac{1}{T^2}\right). \qquad (31)
\end{aligned}$$

Combining (30) and (31), we have that

$$s_T \le \underbrace{\left(\frac{1}{p}(1-\delta)^p e^{-p\delta\log(1/\delta)} + (1-\delta)^p\frac{e^{-p\delta\log(1/\delta)}-1}{1-\delta p}\right)}_{=:s_\infty}\left(1+O\left(\frac{\log T}{T}\right)\right).$$

Also, for all large T's, since $s_\infty \to \frac{1}{p}$ and $\delta\log(1/\delta)\to 0$ as $\delta\to 0$, when $\delta \ll 1$,

$$s_T \le \frac{1}{p} - \left(1+\frac{p+1}{2p}\right)\frac{c\log T}{T} + \tilde{O}\left(\frac{1}{T^2}\right).$$

The proof is now complete. □

I Proofs in Section 4

I.1 Proof of Theorem 3

Fix $t \ge 1$. Using the forward model in (5), we have that $q_{t|0,y}$ is the p.d.f. of $\mathcal{N}(\sqrt{\bar{\alpha}_t}(I_d - H^\dagger H)x_0 + \sqrt{\bar{\alpha}_t}H^\dagger y, \Sigma_{t|0,y})$. Thus,

$$\begin{aligned}
\nabla\log q_{t|y}(x) &= \frac{1}{q_{t|y}(x)}\int_{x_0\in\mathbb{R}^d}\nabla q_{t|0,y}(x|x_0)\mathrm{d}Q_{0|y}(x_0)\\
&= -\frac{1}{q_{t|y}(x)}\Sigma_{t|0,y}^{-1}\int_{x_0\in\mathbb{R}^d} q_{t|0,y}(x|x_0)(x-\sqrt{\bar{\alpha}_t}(I_d-H^\dagger H)x_0 - \sqrt{\bar{\alpha}_t}H^\dagger y)\mathrm{d}Q_{0|y}(x_0)\\
&= -\Sigma_{t|0,y}^{-1}(x-\sqrt{\bar{\alpha}_t}H^\dagger y)\\
&\quad + \frac{\sqrt{\bar{\alpha}_t}}{q_{t|y}(x)}\Sigma_{t|0,y}^{-1}(I_d-H^\dagger H)\int_{x_0\in\mathbb{R}^d} q_{t|0,y}(x|x_0)x_0\mathrm{d}Q_{0|y}(x_0).
\end{aligned}$$

Thus, the equality for $\nabla\log q_{t|y}$ is established because by Lemma 9,

$$(\sigma_y^2 H^\dagger(H^\dagger)^\intercal + (1-\bar{\alpha}_t)I_d)^{-1}(I_d - H^\dagger H) = (1-\bar{\alpha}_t)^{-1}(I_d - H^\dagger H).$$

To see the optimality with $f_{t,y}^*$, fix $t\ge 1$ and $x\in\mathbb{R}^d$. First note that $(I_d - H^\dagger H)f_{t,y}^*(x) = (I_d - H^\dagger H)\Sigma_{t|0,y}^{-1}(\sqrt{\bar{\alpha}_t}H^\dagger y - H^\dagger H x) = 0$ by Lemma 9. Now, suppose that $f_{t,y} = f_{t,y}^* + v$ such that $(I_d - H^\dagger H)f_{t,y} = 0 \implies (I_d - H^\dagger H)v = 0$. From the definition of $\Delta_{t,y}$ in (7),

$$\Delta_{t,y}(x) = (I_d - H^\dagger H)(\nabla\log q_{t|y}(x) - \nabla\log q_t(x)) + (H^\dagger H)\nabla\log q_{t|y}(x) - f_{t,y}(x)$$

where

$$(H^\dagger H)\nabla\log q_{t|y}(x) - f_{t,y}(x)$$

$$
\begin{aligned}
&= (H^\dagger H)\Sigma_{t|0,y}^{-1}(\sqrt{\bar{\alpha}_t}H^\dagger y - x) - \Sigma_{t|0,y}^{-1}(\sqrt{\bar{\alpha}_t}H^\dagger y - H^\dagger H x) - v \\
&= -(H^\dagger H)\Sigma_{t|0,y}^{-1}(I_d - H^\dagger H)x - (I_d - H^\dagger H)\Sigma_{t|0,y}^{-1}(\sqrt{\bar{\alpha}_t}H^\dagger y - H^\dagger H x) - v \\
&= -v
\end{aligned}
$$

where the last line follows from Lemma 9.

Thus, if $v = 0$, then $f_{t,y} = f_{t,y}^*$, and we have

$$
\Delta_{t,y} = (I_d - H^\dagger H)(\nabla \log q_{t|y}(x) - \nabla \log q_t(x)). \tag{32}
$$

Also, if $v \neq 0$, since v is orthogonal to the space induced by $(I_d - H^\dagger H)$, we have

$$
\|\Delta_{t,y}(x)\|^2 = \left\|(I_d - H^\dagger H)(\nabla \log q_{t|y}(x) - \nabla \log q_t(x))\right\|^2 + \|v\|^2 \tag{33}
$$

which is minimized at $v = 0$. The proof is now complete.

I.2 Proof of Theorem 4

Fix $t \geq 2$. Recall that the unconditional score $\nabla \log q_t(x)$ is

$$
\begin{aligned}
\nabla \log q_t(x) &= \frac{1}{q_t(x)} \int_{x_0 \in \mathbb{R}^d} \nabla q_{t|0}(x|x_0) \mathrm{d}Q_0(x_0) \\
&= -\frac{1}{(1-\bar{\alpha}_t)q_t(x)} \int_{x_0 \in \mathbb{R}^d} q_{t|0}(x|x_0)(x - \sqrt{\bar{\alpha}_t}x_0) \mathrm{d}Q_0(x_0) \\
&= -\frac{1}{(1-\bar{\alpha}_t)}x + \frac{\sqrt{\bar{\alpha}_t}}{(1-\bar{\alpha}_t)q_t(x)} \int_{x_0 \in \mathbb{R}^d} q_{t|0}(x|x_0) x_0 \mathrm{d}Q_0(x_0)
\end{aligned}
$$

since $q_{t|0}$ is the p.d.f. of $\mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I_d)$.

In the first half, we consider the case where σ_y^2 is known, and thus $f_{t,y} = f_{t,y}^*$ in (10). Note that from Theorem 3,

$$
\begin{aligned}
\nabla \log q_{t|y}(x) = \Sigma_{t|0,y}^{-1}(\sqrt{\bar{\alpha}_t}H^\dagger y - x) \\
+ \frac{\sqrt{\bar{\alpha}_t}}{q_{t|y}(x)}\Sigma_{t|0,y}^{-1}(I_d - H^\dagger H) \int_{x_0 \in \mathbb{R}^d} q_{t|0,y}(x|x_0) x_0 \mathrm{d}Q_{0|y}(x_0).
\end{aligned}
$$

Here we also recall from Theorem 3 that

$$
\Sigma_{t|0,y} := \bar{\alpha}_t \sigma_y^2 H^\dagger (H^\dagger)^\intercal + (1-\bar{\alpha}_t)I_d.
$$

Since $H^\dagger(H^\dagger)^\intercal$ is positive semi-definite, all its eigenvalues are non-negative. Write the eigen-decomposition as $H^\dagger(H^\dagger)^\intercal = P\mathrm{diag}(D_1, \ldots, D_d)P^\intercal$ where $D_1 \geq \cdots \geq D_d \geq 0$, $\forall i \in [d]$. Then, $\lambda_{min}(\Sigma_{t|0,y}) \geq \bar{\alpha}_t\sigma_y^2 D_d + 1 - \bar{\alpha}_t \geq 1 - \bar{\alpha}_t$, and we get

$$
\left\|\Sigma_{t|0,y}^{-1}\right\| \leq \frac{1}{1-\bar{\alpha}_t}. \tag{34}
$$

Also, from (6), with the $f_{t,y}^*$ in (10),

$$
\begin{aligned}
g_{t,y}(x) &= f_{t,y}^*(x) + (I_d - H^\dagger H)\nabla \log q_t(x) \\
&= \Sigma_{t|0,y}^{-1}\left(\sqrt{\bar{\alpha}_t}H^\dagger y - H^\dagger H x\right) - \frac{1}{(1-\bar{\alpha}_t)}(I_d - H^\dagger H)x \\
&\quad + \frac{\sqrt{\bar{\alpha}_t}}{(1-\bar{\alpha}_t)q_t(x)}(I_d - H^\dagger H) \int_{x_0 \in \mathbb{R}^d} q_{t|0}(x|x_0) x_0 \mathrm{d}Q_0(x_0) \\
&\overset{(i)}{=} \Sigma_{t|0,y}^{-1}\left(\sqrt{\bar{\alpha}_t}H^\dagger y - H^\dagger H x\right) - \Sigma_{t|0,y}^{-1}(I_d - H^\dagger H)x \\
&\quad + \frac{\sqrt{\bar{\alpha}_t}}{q_t(x)}\Sigma_{t|0,y}^{-1}(I_d - H^\dagger H) \int_{x_0 \in \mathbb{R}^d} q_{t|0}(x|x_0) x_0 \mathrm{d}Q_0(x_0) \\
&= \Sigma_{t|0,y}^{-1}\left(\sqrt{\bar{\alpha}_t}H^\dagger y - x\right) + \frac{\sqrt{\bar{\alpha}_t}}{q_t(x)}\Sigma_{t|0,y}^{-1}(I_d - H^\dagger H) \int_{x_0 \in \mathbb{R}^d} q_{t|0}(x|x_0) x_0 \mathrm{d}Q_0(x_0)
\end{aligned}
$$

where (i) follows from Lemma 9. Then, the norm-squared of the score mismatch at time $t \geq 2$ is

$$\begin{aligned}
&\|\Delta_{t,y}\|^2 = \left\|\nabla \log q_{t|y} - g_{t,y}\right\|^2 = \left\|(I_d - H^\dagger H)(\nabla \log q_{t|y} - \nabla \log q_t)\right\|^2 \\
&\leq \bar{\alpha}_t \left\|\Sigma_{t|0,y}^{-1}\right\|^2 \left\| \frac{\int_{x_0 \in \mathbb{R}^d} q_{t|0,y}(x|x_0) x_0 \mathrm{d}Q_{0|y}(x_0)}{q_{t|y}(x)} - \frac{\int_{x_0 \in \mathbb{R}^d} q_{t|0}(x|x_0) x_0 \mathrm{d}Q_0(x_0)}{q_t(x)} \right\|^2 \\
&\overset{(ii)}{\leq} \bar{\alpha}_t \left\|\Sigma_{t|0,y}^{-1}\right\|^2 \int_{x_a, x_b \in \mathbb{R}^d} \|x_a - x_b\|^2 \, \mathrm{d}Q_{0|t,y}(x_a) \mathrm{d}Q_{0|t}(x_b) \\
&\leq \bar{\alpha}_t \left\|\Sigma_{t|0,y}^{-1}\right\|^2 \max_{\substack{x_a \in \mathrm{supp}(Q_{0|y}) \\ x_b \in \mathrm{supp}(Q_0)}} \|x_a - x_b\|^2 \\
&\overset{(iii)}{\lesssim} \frac{\bar{\alpha}_t}{(1-\bar{\alpha}_t)^2} d.
\end{aligned} \tag{35}$$

Here (ii) follows from Jensen's inequality, and (iii) follows by (34) and from the assumption that Q_0 has bounded support (and thus also for both $Q_{0|t}$ and $Q_{0|t,y}$). Therefore, with the α_t in (8) (cf. Lemma 6), since $1 - \bar{\alpha}_t \geq 1 - \delta$ which is a constant, Assumption 4 is satisfied for all $\sigma_y^2 \geq 0$. Thus, Theorem 2 holds with $\gamma = 1$ and $r = 2$.

Now, we consider the case where σ_y^2 is unknown, and the conditional sampler of interest is $g_{t,y}^N(x) = f_{t,y}^N(x) + (I_d - H^\dagger H)\nabla \log q_t(x)$ where $f_{t,y}^N(x) = (1-\bar{\alpha}_t)^{-1}\left(\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H x\right)$. With the same notation as in the proof of Theorem 3, we can write $v = f_{t,y}^N - f_{t,y}^* = ((1-\bar{\alpha}_t)^{-1} I_d - \Sigma_{t|0,y}^{-1})\left(\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H x\right)$. Note that v still satisfies that $(I_d - H^\dagger H)v = 0$. Using the result in (33), we have

$$\begin{aligned}
\left\|\Delta_{t,y}^N\right\|^2 = &\left\|(I_d - H^\dagger H)(\nabla \log q_{t|y}(x) - \nabla \log q_t(x))\right\|^2 \\
&+ \left\|(\Sigma_{t|0,y}^{-1} - (1-\bar{\alpha}_t)^{-1} I_d)\left(\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H x\right)\right\|^2
\end{aligned}$$

where the first term is the same as in (35) which can be upper-bounded in a similar way. To upper-bound the second term, note that by Woodbury matrix identity,

$$\begin{aligned}
\left\|\Sigma_{t|0,y}^{-1} - \frac{1}{1-\bar{\alpha}_t} I_d\right\| &= \frac{\bar{\alpha}_t \sigma_y^2}{(1-\bar{\alpha}_t)^2} \left\| H^\dagger \left(I_p + \frac{\sigma_y^2}{1-\bar{\alpha}_t} (H^\dagger)^\intercal H^\dagger \right)^{-1} (H^\dagger)^\intercal \right\| \\
&\lesssim \frac{\bar{\alpha}_t \sigma_y^2}{(1-\bar{\alpha}_t)^2}
\end{aligned} \tag{36}$$

where the inequality follows because $\left\|H^\dagger\right\| \lesssim 1$ is a constant and the minimum eigenvalue of $(I_p + \frac{\sigma_y^2}{1-\bar{\alpha}_t}(H^\dagger)^\intercal H^\dagger)$ is at least 1. Thus,

$$\begin{aligned}
&\mathbb{E}_{Q_{t|y}} \left\|(\Sigma_{t|0,y}^{-1} - (1-\bar{\alpha}_t)^{-1} I_d)\left(\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H X_t\right)\right\|^2 \\
&\overset{(iv)}{\leq} \frac{\bar{\alpha}_t^2 \sigma_y^4}{(1-\bar{\alpha}_t)^4} \mathbb{E}_{Q_{t|y}} \left\|\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H X_t\right\|^2 \\
&= \frac{\bar{\alpha}_t^2 \sigma_y^4}{(1-\bar{\alpha}_t)^4} \mathbb{E}_{Q_{0|y}} \mathbb{E}_{Q_{t|0,y}} \left\|\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H X_t\right\|^2 \\
&= \frac{\bar{\alpha}_t^2 \sigma_y^4}{(1-\bar{\alpha}_t)^4} \mathbb{E}_{Q_{0|y}} \mathbb{E}_{Q_{t|0}} \left\|\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H X_t\right\|^2 \\
&\leq \frac{2\bar{\alpha}_t^2 \sigma_y^4}{(1-\bar{\alpha}_t)^4} \mathbb{E}_{Q_{0|y}} \left[\bar{\alpha}_t \left\|H^\dagger y - H^\dagger H X_0\right\|^2 + \mathbb{E}_{Q_{t|0}} \left\|H^\dagger H (X_t - \sqrt{\bar{\alpha}_t} X_0)\right\|^2\right] \\
&\leq \frac{2\bar{\alpha}_t^2 \sigma_y^4}{(1-\bar{\alpha}_t)^4} \mathbb{E}_{Q_{0|y}} \left[\bar{\alpha}_t \left\|H^\dagger y\right\|^2 + \bar{\alpha}_t \left\|X_0\right\|^2 + d(1-\bar{\alpha}_t)\right]
\end{aligned}$$

$$\overset{(v)}{\lesssim} \frac{\bar{\alpha}_t^2 \sigma_y^4}{(1-\bar{\alpha}_t)^4} d$$

where (iv) follows from (36), and (v) follows from the fact that $Q_{0|y}$ has bounded support. Similarly, for general moments $\ell \geq 2$,

$$\mathbb{E}_{Q_{t|y}} \left\| (\Sigma_{t|0,y}^{-1} - (1-\bar{\alpha}_t)^{-1} I_d) \left(\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H X_t \right) \right\|^\ell \lesssim \left(\frac{\bar{\alpha}_t \sigma_y^2}{(1-\bar{\alpha}_t)^2} d \right)^{\ell/2}.$$

Therefore, with the α_t in (8), since $1-\bar{\alpha}_t \geq 1-\delta$ which is a constant, we still have that Assumption 4 is satisfied (see Lemma 6), and Theorem 2 still holds with $\gamma = 1$ and $r = 4$. The proof is now complete.

I.3 Theorem 6 and its proof

Before we enter the proof of Proposition 1 and Theorem 5, we first state a similar set of results for Gaussian Q_0, which turns out to be useful for analyzing Gaussian mixture Q_0's. To begin, the following lemma investigates $\mathbb{E}_{X_t \sim Q_{t|y}} \|\Delta_{t,y}(X_t)\|^2$ when Q_0 is Gaussian. This quantity is proportional to the asymptotic bias $\mathcal{W}_{\text{bias}}$.

Proposition 2. *For $Q_0 = \mathcal{N}(\mu_0, \Sigma_0)$, if $f_{t,y} = f_{t,y}^*$ in (10) and $H = (I_p \quad 0)$, with the α_t's according to Definition 1, Assumption 4 is satisfied, and*

$$\begin{aligned}\mathbb{E}_{X_t \sim Q_{t|y}} \|\Delta_{t,y}(X_t)\|^2 &\leq \bar{\alpha}_t^2 \frac{\max\{\left\|H^\dagger y - H^\dagger H \mu_0\right\|^2 + d(\lambda_1 + \sigma_y^2), d\}}{\min\{\lambda_d, 1\}^2 \min\{\tilde{\lambda}_{d-p}, 1\}^2} \|[\Sigma_0]_{y\bar{y}} [\Sigma_0]_{\bar{y}y}\| \\ &\lesssim \bar{\alpha}_t^2 \cdot (\left\|H^\dagger y - H^\dagger H \mu_0\right\|^2 + d)\end{aligned}$$

where λ_1 is the largest eigenvalue of Σ_0, and λ_d and $\tilde{\lambda}_{d-p}$ are the smallest eigenvalues of Σ_0 and $[\Sigma_0]_{\bar{y}\bar{y}}$, respectively.

Proof. See Appendix J.1. □

With this lemma, the following theorem characterizes the conditional KL divergence when Q_0 is Gaussian.

Theorem 6. *Suppose that $\sigma_y^2 > 0$. Suppose that Assumptions 1 and 5 hold. Under the same conditions as in Proposition 2, if α_t further satisfies $\sum_{t=1}^T (1-\alpha_t)\bar{\alpha}_t = 1 + o(1)$, we have*

$$\begin{aligned}\mathrm{KL}(Q_{0|y} \| \widehat{P}_{0|y}) &\lesssim (\left\|H^\dagger y - H^\dagger H \mu_0\right\|^2 + d) \\ &\quad + (\left\|H^\dagger y - H^\dagger H \mu_0\right\|^2 + d) \frac{(\log T)^2}{T} + \sqrt{\left\|H^\dagger y - H^\dagger H \mu_0\right\|^2 + d} \cdot (\log T) \varepsilon.\end{aligned}$$

Note that a similar result can be obtained for $\mathrm{KL}(Q_{1|y} \| \widehat{P}_{1|y})$ (where $\mathrm{W}_2(Q_{1|y}, Q_{0|y})^2 \lesssim \delta d$) for any general $\sigma_y^2 \geq 0$ using the α_t in (8) (see Remark 1).

I.3.1 Proof of Theorem 6

Throughout the proof we use the same notations as in (55). Since Assumption 4 is satisfied from Proposition 2, in order to invoke Theorem 1, we still need to check Assumption 3. Since each $Q_{t|y}$ ($\forall t \geq 0$) is Gaussian, all partial derivatives of its log-p.d.f. higher than third-order equal zero. For the first and second-order, note that $\Sigma_{t|y} = \bar{\alpha}_t (I_d - H^\dagger H) \Sigma_0 (I_d - H^\dagger H) + (1 - \bar{\alpha}_t) I_d + \bar{\alpha}_t \sigma_y^2 H^\dagger H$. Thus, when $\sigma_y^2 > 0$, $\lambda_{min}(\Sigma_{t-1|y}) \geq \min\{1 - \bar{\alpha}_t + \bar{\alpha}_t \sigma_y^2, 1 - \bar{\alpha}_t + \bar{\alpha}_t \tilde{\lambda}_d\} \geq \min\{1, \sigma_y^2, \tilde{\lambda}_d\} > 0$, which yields

$$\left\| \Sigma_{t-1|y}^{-1} \right\| \lesssim 1, \quad \forall t \geq 1. \tag{37}$$

Thus, we have, $\forall \ell \geq 1$,

$$\mathbb{E}_{Q_{t|y}} \left\| \nabla \log q_{t|y}(X_t) \right\|^\ell$$

$$
\begin{aligned}
&= \mathbb{E}_{Q_{t|y}} \left\| \Sigma_{t|y}^{-1}(X_t - \mu_{t|y}) \right\|^\ell \le \left\| \Sigma_{t|y}^{-\frac{1}{2}} \right\|^\ell \mathbb{E}_{Q_{t|y}} \left\| \Sigma_{t|y}^{-\frac{1}{2}}(X_t - \mu_{t|y}) \right\|^\ell \\
&\lesssim d^{\ell/2} = O(1), \\
&\mathbb{E}_{Q_{t|y}} \left\| \nabla \log q_{t-1|y}(m_{t,y}(X_t)) \right\|^\ell \\
&= \mathbb{E}_{Q_{t|y}} \left\| \Sigma_{t|y}^{-1}(m_{t,y}(X_t) - \mu_{t|y}) \right\|^\ell \\
&\lesssim \left\| \Sigma_{t|y}^{-\frac{1}{2}} \right\|^\ell \mathbb{E}_{Q_{t|y}} \left\| \Sigma_{t|y}^{-\frac{1}{2}}(X_t - \mu_{t|y}) \right\|^\ell + \left\| \Sigma_{t|y}^{-1} \right\|^\ell \mathbb{E}_{Q_{t|y}} \left\| \nabla \log q_{t|y}(X_t) \right\|^\ell \\
&\lesssim d^{\ell/2} = O(1), \\
&\mathbb{E}_{Q_{t|y}} \left\| \nabla^2 \log q_{t|y}(X_t) \right\|^\ell = \left\| \Sigma_{t|y}^{-1} \right\|^\ell = O(1), \\
&\mathbb{E}_{Q_{t|y}} \left\| \nabla^2 \log q_{t-1|y}(m_{t,y}(X_t)) \right\|^\ell = \left\| \Sigma_{t-1|y}^{-1} \right\|^\ell = O(1).
\end{aligned}
$$

Thus, Assumption 3 holds when $1 - \alpha_t$ satisfies Definition 1.

Now, we can invoke Theorem 1 and get $\mathrm{KL}(Q_{0|y} \| \widehat{P}_{0|y}) \lesssim \mathcal{W}_{\text{oracle}} + \mathcal{W}_{\text{bias}} + \mathcal{W}_{\text{vanish}}$, where

$$
\begin{aligned}
\mathcal{W}_{\text{oracle}} &= \sum_{t=1}^{T} \frac{(1-\alpha_t)^2}{2\alpha_t} \mathbb{E}_{X_t \sim Q_{t|y}} \left[\mathrm{Tr}\Big(\nabla^2 \log q_{t-1|y}(m_{t,y}(X_t)) \nabla^2 \log q_{t|y}(X_t) \Big) \right] + (\log T)\varepsilon^2 \\
\mathcal{W}_{\text{bias}} &= \sum_{t=1}^{T} (1-\alpha_t) \mathbb{E}_{X_t \sim Q_{t|y}} \left\| \Delta_{t,y}(X_t) \right\|^2 \\
\mathcal{W}_{\text{vanish}} &= \sum_{t=1}^{T} \frac{1-\alpha_t}{\sqrt{\alpha_t}} \mathbb{E}_{X_t \sim Q_{t|y}} \left[(\nabla \log q_{t-1|y}(m_{t,y}(X_t)) - \sqrt{\alpha_t} \nabla \log q_{t|y}(X_t))^\intercal \Delta_{t,y}(X_t) \right] \\
&\quad - \sum_{t=1}^{T} \frac{(1-\alpha_t)^2}{2\alpha_t} \mathbb{E}_{X_t \sim Q_{t|y}} \left[\Delta_{t,y}(X_t)^\intercal \nabla^2 \log q_{t-1|y}(m_{t,y}(X_t)) \Delta_{t,y}(X_t) \right] \\
&\quad + \sum_{t=1}^{T} \frac{(1-\alpha_t)^2}{3! {\alpha_t}^{3/2}} \mathbb{E}_{X_t \sim Q_{t|y}} \Bigg[3 \sum_{i=1}^{d} \partial_{iii}^3 \log q_{t-1|y}(m_{t,y}(X_t)) \Delta_{t,y}(X_t)^i \\
&\qquad + \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \partial_{iij}^3 \log q_{t-1|y}(m_{t,y}(X_t)) \Delta_{t,y}(X_t)^j \Bigg] \\
&\quad + \max_{t \ge 1} \sqrt{\mathbb{E}_{X_t \sim Q_{t|y}} \left\| \Delta_{t,y}(X_t) \right\|^2} (\log T)\varepsilon.
\end{aligned}
$$

We first consider the estimation error (in both $\mathcal{W}_{\text{oracle}}$ and $\mathcal{W}_{\text{vanish}}$), which can be upper-bounded as

$$
\max_{t \ge 1} \sqrt{\mathbb{E}_{X_t \sim Q_{t|y}} \left\| \Delta_{t,y}(X_t) \right\|^2} (\log T)\varepsilon + (\log T)\varepsilon^2 \lesssim \left(\left\| H^\dagger y - H^\dagger H \mu_0 \right\|^2 + d \right)^{\frac{1}{2}} (\log T)\varepsilon
$$

from Proposition 2. Also, when $Q_{t|y}$ is Gaussian, we can calculate, for any $x_t \in \mathbb{R}^d$,

$$
\begin{aligned}
&\mathrm{Tr}\Big(\nabla^2 \log q_{t-1|y}(m_{t,y}(x_t)) \nabla^2 \log q_{t|y}(x_t) \Big) = \mathrm{Tr}(\Sigma_{t-1|y}^{-1} \Sigma_{t|y}^{-1}) \\
&= \mathrm{Tr}(\Sigma_{t-1|y}^{-1} (\alpha_t \Sigma_{t-1|y} + (1-\alpha_t) I_d)^{-1}) \\
&\overset{(i)}{=} \mathrm{Tr}(\Sigma_{t-1|y}^{-1} (\alpha_t^{-1} \Sigma_{t-1|y}^{-1} - \frac{1-\alpha_t}{\alpha_t^2} \Sigma_{t-1|y}^{-2} + O((1-\alpha_t)^2))) \\
&\lesssim \frac{1}{\alpha_t} \mathrm{Tr}(\Sigma_{t-1|y}^{-2})
\end{aligned}
$$

where (i) follows from Taylor expansion when $1 - \alpha_t$ is small. Using (37), this implies that

$$
\mathcal{W}_{\text{oracle}} \lesssim \frac{d(\log T)^2}{T} + (\log T)\varepsilon^2.
$$

Also, from the condition on α_t,

$$\sum_{t=1}^{T}(1-\alpha_t)\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}(X_t)\right\|^2 \lesssim (\left\|H^\dagger y - H^\dagger H\mu_0\right\|^2 + d).$$

Now we focus on $\mathcal{W}_{\text{vanish}}$ (except the estimation error). Since $Q_{t|y}$ is Gaussian, all third-order partial derivatives are zero, and only the first two terms in $\mathcal{W}_{\text{vanish}}$ remain. In the following we fix $t \geq 1$. Also recall from (56) that when $H = (I_p \quad 0)$,

$$\begin{aligned}\Delta_{t,y} &= -\bar{\alpha}_t(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H)\Sigma_0(H^\dagger H)\Sigma_t^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_0)\\ &= -\bar{\alpha}_t(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}[\Sigma_0]_{\bar{y}y}[\Sigma_t^{-1}]_{y:}(x_t - \sqrt{\bar{\alpha}_t}\mu_0).\end{aligned}$$

For the first term of $\mathcal{W}_{\text{vanish}}$, we first calculate for each x_t that

$$\begin{aligned}&\nabla \log q_{t-1|y}(m_{t,y}) - \sqrt{\alpha_t}\nabla \log q_{t|y}(x_t)\\ &= \sqrt{\alpha_t}\Sigma_{t|y}^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_{0|y}) - \Sigma_{t-1|y}^{-1}(m_{t,y} - \sqrt{\bar{\alpha}_{t-1}}\mu_{0|y})\end{aligned}$$

Recall that

$$\begin{aligned}m_{t,y} &= \frac{1}{\sqrt{\alpha_t}}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla \log q_{t|y}(x_t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\Sigma_{t|y}^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_{0|y}),\\ \Sigma_{t|y}^{-1} &= (\alpha_t\Sigma_{t-1|y} + (1-\alpha_t)I_d)^{-1} = \frac{1}{\alpha_t}\Sigma_{t-1|y}^{-1} - \frac{1-\alpha_t}{\alpha_t^2}\Sigma_{t-1|y}^{-2} + O((1-\alpha_t)^2).\end{aligned}$$

Thus,

$$\begin{aligned}&\nabla \log q_{t-1|y}(m_{t,y}) - \sqrt{\alpha_t}\nabla \log q_{t|y}(x_t)\\ &= \sqrt{\alpha_t}\left(\frac{1}{\alpha_t}\Sigma_{t-1|y}^{-1} - \frac{1-\alpha_t}{\alpha_t^2}\Sigma_{t-1|y}^{-2}\right)(x_t - \sqrt{\bar{\alpha}_t}\mu_{0|y})\\ &\quad - \Sigma_{t-1|y}^{-1}\left(\frac{1}{\sqrt{\alpha_t}}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla \log q_{t|y}(x_t) - \sqrt{\bar{\alpha}_{t-1}}\mu_{0|y}\right) + O((1-\alpha_t)^2)\\ &= -\frac{1-\alpha_t}{\alpha_t^{3/2}}\Sigma_{t-1|y}^{-2}(x_t - \sqrt{\bar{\alpha}_t}\mu_{0|y}) + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\Sigma_{t-1|y}^{-1}\Sigma_{t|y}^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_{0|y}) + O((1-\alpha_t)^2).\end{aligned}$$

Combining with the definition for $\Delta_{t,y}$ in (56) and using Lemma 9, we have

$$\begin{aligned}&\mathbb{E}_{X_t\sim Q_{t|y}}\left[\Delta_{t,y}(X_t)^\intercal(\nabla \log q_{t-1|y}(m_{t,y}(X_t)) - \sqrt{\alpha_t}\nabla \log q_{t|y}(X_t))\right]\\ &= \bar{\alpha}_t\mathbb{E}_{X_t\sim Q_{t|y}}\Bigg[(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal\Sigma_t^{-1}(H^\dagger H)\Sigma_0(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H)\\ &\quad \left(\frac{1-\alpha_t}{\alpha_t^{3/2}}\Sigma_{t-1|y}^{-2} - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\Sigma_{t-1|y}^{-1}\Sigma_{t|y}^{-1}\right)(X_t - \sqrt{\bar{\alpha}_t}\mu_{0|y})\Bigg] + O((1-\alpha_t)^2)\\ &\overset{(ii)}{=} \bar{\alpha}_t\mathbb{E}_{X_t\sim Q_{t|y}}\Bigg[(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal\Sigma_t^{-1}(H^\dagger H)\Sigma_0(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H)\\ &\quad \left(\frac{1-\alpha_t}{\alpha_t^{3/2}}\Sigma_{t-1|y}^{-1}(I_d - H^\dagger H)\Sigma_{t-1|y}^{-1} - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\Sigma_{t-1|y}^{-1}(I_d - H^\dagger H)\Sigma_{t|y}^{-1}\right)\\ &\quad (I_d - H^\dagger H)(X_t - \sqrt{\bar{\alpha}_t}\mu_0)\Bigg] + O((1-\alpha_t)^2)\\ &= \bar{\alpha}_t\mathbb{E}_{X_t\sim Q_{t|y}}\Bigg[(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal[\Sigma_t^{-1}]_{:y}[\Sigma_0]_{y\bar{y}}(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}\\ &\quad \left(\frac{1-\alpha_t}{\alpha_t^{3/2}}[\Sigma_{t-1|y}^{-1}]_{\bar{y}\bar{y}}^2 - \frac{1-\alpha_t}{\sqrt{\alpha_t}}[\Sigma_{t-1|y}^{-1}]_{\bar{y}\bar{y}}[\Sigma_{t|y}^{-1}]_{\bar{y}\bar{y}}\right)(0 \quad I_{d-p})(X_t - \sqrt{\bar{\alpha}_t}\mu_0)\Bigg]\\ &\quad + O((1-\alpha_t)^2)\end{aligned}$$

$$
\begin{aligned}
&= \bar{\alpha}_t \mathrm{Tr}\Bigg([\Sigma_t^{-1}]_{:y} [\Sigma_0]_{y\bar{y}} (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} \left(\frac{1-\alpha_t}{\alpha_t^{3/2}} [\Sigma_{t-1|y}^{-1}]^2_{\bar{y}\bar{y}} - \frac{1-\alpha_t}{\sqrt{\alpha_t}} [\Sigma_{t-1|y}^{-1}]_{\bar{y}\bar{y}} [\Sigma_{t|y}^{-1}]_{\bar{y}\bar{y}} \right) \\
&\quad (0 \quad I_{d-p}) \, \mathbb{E}_{X_t \sim Q_{t|y}} \left[(X_t - \sqrt{\bar{\alpha}_t}\mu_0)(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal \right] \Bigg) \\
&\quad + O((1-\alpha_t)^2) \\
&\overset{(iii)}{\leq} \left\| [\Sigma_t^{-1}]_{:y} \right\| \left\| [\Sigma_0]_{y\bar{y}} \right\| \left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} \right\| \times \\
&\quad \left(\frac{1-\alpha_t}{\alpha_t^{3/2}} \left\| [\Sigma_{t-1|y}^{-1}]_{\bar{y}\bar{y}} \right\|^2 + \frac{1-\alpha_t}{\sqrt{\alpha_t}} \left\| [\Sigma_{t-1|y}^{-1}]_{\bar{y}\bar{y}} \right\| \left\| [\Sigma_{t|y}^{-1}]_{\bar{y}\bar{y}} \right\| \right) \times \\
&\quad \mathrm{Tr}\left(\mathbb{E}_{X_t \sim Q_{t|y}} \left[(X_t - \sqrt{\bar{\alpha}_t}\mu_0)(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal \right] \right) \\
&\overset{(iv)}{\lesssim} \left(\frac{1-\alpha_t}{\alpha_t^{3/2}} + \frac{1-\alpha_t}{\sqrt{\alpha_t}} \right) \max\left\{ \left\| H^\dagger y - H^\dagger H \mu_0 \right\|^2 + d(\lambda_1 + \sigma_y^2), d \right\} \\
&\lesssim (1-\alpha_t) \left(\left\| H^\dagger y - H^\dagger H \mu_0 \right\|^2 + d \right)
\end{aligned}
$$

where (ii) follows by Lemma 9 and from definition that $(I_d - H^\dagger H)\mu_{0|y} = (I_d - H^\dagger H)\mu_0$, (iii) follows because $|\mathrm{Tr}(UV)| \leq \|U\| \, \mathrm{Tr}(V)$ if V is positive semi-definite, and (iv) follows from (58) and the same reasons for (57). In particular, we note that $[\Sigma_{t|y}^{-1}]_{\bar{y}\bar{y}} = [\Sigma_{t,sig}^{-1}]_{\bar{y}\bar{y}} = (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1}$ and $\left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} \right\| \leq \frac{1}{\min\{\tilde{\lambda}_{d-p}, 1\}} < \infty$.

For the second term of $\mathcal{W}_{\text{vanish}}$, we use the fact that $[\Sigma_{t-1|y}^{-1}]_{\bar{y}\bar{y}} = (\bar{\alpha}_{t-1} [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_{t-1}) I_{d-p})^{-1}$ and have

$$
\begin{aligned}
&- \mathbb{E}_{X_t \sim Q_{t|y}} \Delta_{t,y}(X_t)^\intercal \nabla^2 \log q_{t-1|y}(m_{t,y}(X_t)) \Delta_{t,y}(X_t) \\
&= \bar{\alpha}_t^2 \mathbb{E}_{X_t \sim Q_{t|y}} \Bigg[(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal \Sigma_t^{-1} (H^\dagger H) \Sigma_0 (I_d - H^\dagger H) \Sigma_{t,sig}^{-1} (I_d - H^\dagger H) \Sigma_{t-1|y}^{-1} \\
&\quad (I_d - H^\dagger H) \Sigma_{t,sig}^{-1} (I_d - H^\dagger H) \Sigma_0 (H^\dagger H) \Sigma_t^{-1} (X_t - \sqrt{\bar{\alpha}_t}\mu_0) \Bigg] \\
&= \bar{\alpha}_t^2 \mathrm{Tr}\Bigg([\Sigma_t^{-1}]_{:y} [\Sigma_0]_{y\bar{y}} (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} (\bar{\alpha}_{t-1} [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_{t-1}) I_{d-p})^{-1} \\
&\quad (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} [\Sigma_0]_{\bar{y}y} [\Sigma_t^{-1}]_{y:} \mathbb{E}_{X_t \sim Q_{t|y}} (X_t - \sqrt{\bar{\alpha}_t}\mu_0)(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal \Bigg) \\
&\overset{(v)}{\leq} \max\left\{ \left\| H^\dagger y - H^\dagger H \mu_0 \right\|^2 + d(\lambda_1 + \sigma_y^2), d \right\} \left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} \right\|^2 \\
&\quad \left\| (\bar{\alpha}_{t-1} [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_{t-1}) I_{d-p})^{-1} \right\| \left\| \Sigma_t^{-1} \right\|^2 \left\| [\Sigma_0]_{y\bar{y}} [\Sigma_0]_{\bar{y}y} \right\| \\
&\lesssim \max\left\{ \left\| H^\dagger y - H^\dagger H \mu_0 \right\|^2 + d(\lambda_1 + \sigma_y^2), d \right\} \\
&\lesssim \left\| H^\dagger y - H^\dagger H \mu_0 \right\|^2 + d.
\end{aligned}
$$

Here (v) follows from the fact that $|\mathrm{Tr}(UV)| \leq \|U\| \, \mathrm{Tr}(V)$ if V is positive semi-definite and (58). The proof is complete by plugging all the results above into Theorem 1.

Remark 1. Before we end the proof, we leave a note for the case of $\sigma_y^2 = 0$ (indeed, for any general $\sigma_y^2 \geq 0$). The only difference is how to upper-bound $\mathcal{W}_{\text{oracle}}$. In particular, if $\sigma_y^2 = 0$, (37) no longer holds (i.e., we can no longer upper-bound $\left\| \Sigma_{t-1|y}^{-1} \right\|$ as a constant). Instead, we can obtain an upper bound as $\left\| \Sigma_{t-1|y}^{-1} \right\| \lesssim (1-\bar{\alpha}_{t-1})^{-1}$. Then, with the α_t in (8), we have

$$
\mathcal{W}_{\text{oracle}} \lesssim \frac{d(\log T)^2 \log(1/\delta)^2}{T} + (\log T)\varepsilon^2.
$$

The rest of the proof still follows because the α_t satisfies Definition 1 when $t \geq 2$. Combining with Lemma 7, we would finally obtain

$$\mathrm{KL}(Q_{1|y}\|\widehat{P}_{1|y}) \lesssim (\left\|H^\dagger y - H^\dagger H\mu_0\right\|^2 + d)\left(1 - \frac{3\log(1/\delta)\log T}{T}\right)$$
$$+ (\left\|H^\dagger y - H^\dagger H\mu_0\right\|^2 + d)\frac{(\log T)^2\log(1/\delta)^2}{T} + \sqrt{\|H^\dagger y - H^\dagger H\mu_0\|^2 + d}\cdot(\log T)\varepsilon.$$

Here $\mathrm{W}_2(Q_{1|y}, Q_{0|y})^2 \lesssim \delta d$.

I.4 Proof of Proposition 1

We first introduce some useful notations for this subsection. Recall that Q_0 has mixture p.d.f. in which the mixture prior π_n is independent of y ($= Hx_0 + n$). Thus, using the fact that $x_0 = (I_d - H^\dagger H)x_0 + H^\dagger y - H^\dagger n$, we can define $Q_{0,n|y}$ as (cf. Flåm (2013))

$$Q_{0|y} = \sum_{n=1}^N \pi_n Q_{0,n|y}$$
$$:= \sum_{n=1}^N \pi_n \mathcal{N}((I_d - H^\dagger H)\mu_{0,n} + H^\dagger y, (I_d - H^\dagger H)\Sigma_0(I_d - H^\dagger H) + \sigma_y^2 H^\dagger (H^\dagger)^\intercal).$$

Note that when $H = (I_p \quad 0)$ and $\sigma_y^2 > 0$, $q_{0|y}$ exists. From the conditional forward model in (5), we further define

$$Q_t = \sum_{n=1}^N \pi_n Q_{t,n},\ \ Q_{t|y} = \sum_{n=1}^N \pi_n Q_{t,n|y},\ \ Q_{t,n} := \mathcal{N}(\mu_{t,n}, \Sigma_t),\ \ Q_{t,n|y} := \mathcal{N}(\mu_{t,n|y}, \Sigma_{t|y})$$
$$\mu_{t,n} := \sqrt{\bar\alpha_t}\mu_{0,n},\ \ \Sigma_t := \bar\alpha_t\Sigma_0 + (1-\bar\alpha_t)I_d,\ \ \mu_{t,n|y} := \sqrt{\bar\alpha_t}(I_d - H^\dagger H)\mu_{0,n} + \sqrt{\bar\alpha_t}H^\dagger y$$
$$\Sigma_{t|y} := \Sigma_{t,sig} + \bar\alpha_t\sigma_y^2 H^\dagger(H^\dagger)^\intercal,\ \ \Sigma_{t,sig} := \bar\alpha_t(I_d - H^\dagger H)\Sigma_0(I_d - H^\dagger H) + (1-\bar\alpha_t)I_d. \quad (38)$$

Similar to (37), we still have

$$\left\|\Sigma_{t-1|y}^{-1}\right\| \lesssim 1, \quad \forall t \geq 1.$$

We can also calculate the scores of Q_t and $Q_{t|y}$ in as follows.

$$\nabla \log q_t(x_t) = -\frac{1}{q_t(x_t)}\sum_{n=1}^N \pi_n q_{t,n}(x_t)\Sigma_t^{-1}(x_t - \mu_{t,n}),$$
$$\nabla \log q_{t|y}(x_t) = -\frac{1}{q_{t|y}(x_t)}\sum_{n=1}^N \pi_n q_{t,n|y}(x_t)\Sigma_{t|y}^{-1}(x_t - \mu_{t,n|y}). \quad (39)$$

Now, with $f_{t,y} = f_{t,y}^*$ (in (10)), from the expression of $\Delta_{t,y}$ in (32), under the assumption $H = (I_p \quad 0)$, the score mismatch at each diffusion step is equal to

$$\Delta_{t,y} = (I_d - H^\dagger H)(\nabla \log q_{t|y} - \nabla \log q_t)$$
$$= \frac{1}{q_t(x_t)}\sum_{n=1}^N \pi_n q_{t,n}(x_t)(I_d - H^\dagger H)\Sigma_t^{-1}(x_t - \sqrt{\bar\alpha_t}\mu_{0,n})$$
$$- \frac{1}{q_{t|y}(x_t)}\sum_{n=1}^N \pi_n q_{t,n|y}(x_t)(I_d - H^\dagger H)\Sigma_{t|y}^{-1}(x_t - \sqrt{\bar\alpha_t}\mu_{0,n|y})$$
$$= \sum_{n=1}^N \pi_n \left(\frac{q_{t,n}(x_t)}{q_t(x_t)} - \frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}\right)(I_d - H^\dagger H)\Sigma_t^{-1}(x_t - \sqrt{\bar\alpha_t}\mu_{0,n})$$
$$+ \sum_{n=1}^N \pi_n \frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}(I_d - H^\dagger H)\left(\Sigma_t^{-1}(x_t - \sqrt{\bar\alpha_t}\mu_{0,n}) - \Sigma_{t|y}^{-1}(x_t - \sqrt{\bar\alpha_t}\mu_{0,n|y})\right)$$

$$
\begin{aligned}
&= -\sqrt{\bar{\alpha}_t}\sum_{n=1}^{N}\pi_n\left(\frac{q_{t,n}(x_t)}{q_t(x_t)}-\frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}\right)(I_d-H^\dagger H)\Sigma_t^{-1}\mu_{0,n}\\
&\quad+\sum_{n=1}^{N}\pi_n\frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}(I_d-H^\dagger H)\left(\Sigma_t^{-1}(x_t-\sqrt{\bar{\alpha}_t}\mu_{0,n})-\Sigma_{t|y}^{-1}(x_t-\sqrt{\bar{\alpha}_t}\mu_{0,n|y})\right)\\
&\overset{(i)}{=} -\sqrt{\bar{\alpha}_t}\sum_{n=1}^{N}\pi_n\left(\frac{q_{t,n}(x_t)}{q_t(x_t)}-\frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}\right)(I_d-H^\dagger H)\Sigma_t^{-1}\mu_{0,n}\\
&\quad-\bar{\alpha}_t\sum_{n=1}^{N}\pi_n\frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}A_t\Sigma_0(H^\dagger H)\Sigma_t^{-1}(x_t-\sqrt{\bar{\alpha}_t}\mu_{0,n})
\end{aligned}
\tag{40}
$$

where $A_t := (I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H)$. Here (i) follows from similar arguments as in (56). Note that since $H = (I_p \quad 0)$, we have equivalently $A_t = (I_d - H^\dagger H)\Sigma_{t|y}^{-1}(I_d - H^\dagger H)$. Since $H^\dagger H = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}$, we can also re-express the second term in $\Delta_{t,y}$ such that $[\Delta_{t,y}]_y = 0$ and

$$
\begin{aligned}
[\Delta_{t,y}]_{\bar{y}} &= -\sqrt{\bar{\alpha}_t}\sum_{n=1}^{N}\pi_n\left(\frac{q_{t,n}(x_t)}{q_t(x_t)}-\frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}\right)[\Sigma_t^{-1}]_{\bar{y}:}\mu_{0,n}\\
&\quad-\bar{\alpha}_t\sum_{n=1}^{N}\pi_n\frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}}+(1-\bar{\alpha}_t)I_{d-p})^{-1}[\Sigma_0]_{\bar{y}y}[\Sigma_t^{-1}]_{y:}(x_t-\sqrt{\bar{\alpha}_t}\mu_{0,n})
\end{aligned}
\tag{41}
$$

since when $H^\dagger H = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}$, $A_t = \begin{pmatrix} 0 & 0 \\ 0 & (\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}}+(1-\bar{\alpha}_t)I_{d-p})^{-1} \end{pmatrix}$.

Now, for the second moment, we follow similar analyses in (57) and get

$$
\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}\right\|^2\\
&\le 4\bar{\alpha}_t\mathbb{E}_{X_t\sim Q_{t|y}}\max_{n\in[N]}\left\|\Sigma_t^{-1}\mu_{0,n}\right\|^2\\
&\quad+2\bar{\alpha}_t^2\left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}}+(1-\bar{\alpha}_t)I_{d-p})^{-1}\right\|^2\left\|[\Sigma_t^{-1}]_{y:}\right\|^2\left\|[\Sigma_0]_{y\bar{y}}\right\|^2\times\\
&\qquad\mathbb{E}_{X_t\sim Q_{t|y}}\left[\sum_{n=1}^{N}\pi_n\frac{q_{t,n|y}(X_t)}{q_{t|y}(X_t)}\left\|X_t-\sqrt{\bar{\alpha}_t}\mu_{0,n}\right\|^2\right]
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}}\left[\sum_{n=1}^{N}\pi_n\frac{q_{t,n|y}(X_t)}{q_{t|y}(X_t)}\left\|X_t-\sqrt{\bar{\alpha}_t}\mu_{0,n}\right\|^2\right]\\
&=\mathbb{E}_{X_t\sim Q_{t|y}}\mathbb{E}_{N\sim\Pi_{\cdot|t,y}}\left\|X_t-\sqrt{\bar{\alpha}_t}\mu_{0,N}\right\|^2\\
&=\mathbb{E}_{N\sim\Pi_{\cdot|y}}\mathbb{E}_{X_t\sim Q_{t,N|y}}\left\|X_t-\sqrt{\bar{\alpha}_t}\mu_{0,N}\right\|^2\\
&\overset{(ii)}{=}\mathbb{E}_{N\sim\Pi_{\cdot|y}}\left[\mathrm{Tr}(\Sigma_{t|y})+\bar{\alpha}_t\left\|H^\dagger y-H^\dagger H\mu_{0,N}\right\|^2\right]\\
&=\mathrm{Tr}(\Sigma_{t|y})+\bar{\alpha}_t\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^2
\end{aligned}
$$

where (ii) follows from (58) and note that $Q_{t,N|y}$ is Gaussian for each $N = n$. Denote $\lambda_1 \ge \cdots \ge \lambda_d > 0$ and $\tilde{\lambda}_1 \ge \cdots \ge \tilde{\lambda}_{d-p} > 0$ to be the eigenvalues of Σ_0 and $[\Sigma_0]_{\bar{y}\bar{y}}$, respectively. Similarly as the proof of Proposition 2, we have $\|[\Sigma_0]_{\bar{y}y}\| \le \|\Sigma_0\| = \lambda_1$, $\left\|[\Sigma_t^{-1}]_{y:}\right\| \le \left\|\Sigma_t^{-1}\right\| \le (\bar{\alpha}_t\lambda_d + (1-\bar{\alpha}_t))^{-1} \le \frac{1}{\min\{\lambda_d,1\}}$, $\left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}}+(1-\bar{\alpha}_t)I_{d-p})^{-1}\right\| \le \frac{1}{\bar{\alpha}_t\tilde{\lambda}_{d-p}+(1-\bar{\alpha}_t)} \le \frac{1}{\min\{\tilde{\lambda}_{d-p},1\}}$, and $\left\|\Sigma_{t|y}\right\| \le \bar{\alpha}_t(\lambda_1+\sigma_y^2)+(1-\bar{\alpha}_t)$. Therefore,

$\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}\right\|^2$

$$\lesssim \bar{\alpha}_t d + \bar{\alpha}_t^2 \frac{\left\|[\Sigma_0]_{y\bar{y}}\right\|^2}{\min\{\tilde{\lambda}_{d-p},1\}^2 \min\{\lambda_d,1\}^2} \times$$
$$\left(d(1-\bar{\alpha}_t) + \bar{\alpha}_t d(\lambda_1+\sigma_y^2) + \bar{\alpha}_t \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2 \right)$$
$$\lesssim \bar{\alpha}_t d + \bar{\alpha}_t^2 \frac{\left\|[\Sigma_0]_{y\bar{y}}\right\|^2}{\min\{\tilde{\lambda}_{d-p},1\}^2 \min\{\lambda_d,1\}^2} \max\left\{ d(\lambda_1+\sigma_y^2) + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2, d \right\}.$$

The proof is complete.

I.5 Proof of Theorem 5

We first recall all the notations in (38) under Gaussian mixture. We also recall the scores from (39):

$$\begin{aligned}
\nabla \log q_t(x_t) &= -\frac{1}{q_t(x_t)} \sum_{n=1}^N \pi_n q_{t,n}(x_t) \Sigma_t^{-1}(x_t - \mu_{t,n}) \\
&= -\Sigma_t^{-1} x_t + \frac{1}{q_t(x_t)} \sum_{n=1}^N \pi_n q_{t,n}(x_t) \Sigma_t^{-1} \mu_{t,n} \\
\nabla \log q_{t|y}(x_t) &= -\frac{1}{q_{t|y}(x_t)} \sum_{n=1}^N \pi_n q_{t,n|y}(x_t) \Sigma_{t|y}^{-1}(x_t - \mu_{t,n|y}) \\
&= -\Sigma_{t|y}^{-1} x_t + \frac{1}{q_{t|y}(x_t)} \sum_{n=1}^N \pi_n q_{t,n|y}(x_t) \Sigma_{t|y}^{-1} \mu_{t,n|y}
\end{aligned}$$

Also, we recall the explicit expression of $\Delta_{t,y}$ from (41), such that $[\Delta_{t,y}]_y = 0$ and

$$\begin{aligned}
[\Delta_{t,y}]_{\bar{y}} = &-\sqrt{\bar{\alpha}_t} \sum_{n=1}^N \pi_n \left(\frac{q_{t,n}(x_t)}{q_t(x_t)} - \frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)} \right) [\Sigma_t^{-1}]_{\bar{y}:} \mu_{0,n} \\
&- \bar{\alpha}_t \sum_{n=1}^N \pi_n \frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)} (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} [\Sigma_0]_{\bar{y}y} [\Sigma_t^{-1}]_{y:} (x_t - \sqrt{\bar{\alpha}_t}\mu_{0,n}).
\end{aligned}$$

In order to invoke Theorem 1, we need to check Assumptions 3 and 4. From (Liang et al., 2024, Lemmas 13 and 14), since $\left\|\Sigma_{t|y}^{-1}\right\| \lesssim 1$ for all $t \geq 0$, the absolute values of any-order partial derivative are bounded by $O(1)$ in expectation, and thus Assumption 3 is satisfied. The following lemma verifies Assumption 4 using the α_t in Definition 1.

Lemma 8. *Under the same condition of Theorem 5, Assumption 4 holds if the α_t satisfies Definition 1.*

Proof. See Appendix J.2. □

Now we start to upper-bound the conditional KL-divergence of interest. Recall that from Theorem 1, $\mathrm{KL}(Q_{0|y} \| \widehat{P}_{0|y}) \lesssim \mathcal{W}_{\text{oracle}} + \mathcal{W}_{\text{bias}} + \mathcal{W}_{\text{vanish}}$, where

$$\begin{aligned}
\mathcal{W}_{\text{oracle}} &= \sum_{t=1}^T \frac{(1-\alpha_t)^2}{2\alpha_t} \mathbb{E}_{X_t \sim Q_{t|y}} \left[\mathrm{Tr}\Big(\nabla^2 \log q_{t-1|y}(m_{t,y}(X_t)) \nabla^2 \log q_{t|y}(X_t) \Big) \right] + (\log T)\varepsilon^2 \\
\mathcal{W}_{\text{bias}} &= \sum_{t=1}^T (1-\alpha_t) \mathbb{E}_{X_t \sim Q_{t|y}} \left\| \Delta_{t,y}(X_t) \right\|^2 \\
\mathcal{W}_{\text{vanish}} &= \sum_{t=1}^T \frac{1-\alpha_t}{\sqrt{\alpha_t}} \mathbb{E}_{X_t \sim Q_{t|y}} \left[(\nabla \log q_{t-1|y}(m_{t,y}(X_t)) - \sqrt{\alpha_t} \nabla \log q_{t|y}(X_t))^\intercal \Delta_{t,y}(X_t) \right] \\
&\quad - \sum_{t=1}^T \frac{(1-\alpha_t)^2}{2\alpha_t} \mathbb{E}_{X_t \sim Q_{t|y}} \left[\Delta_{t,y}(X_t)^\intercal \nabla^2 \log q_{t-1|y}(m_{t,y}(X_t)) \Delta_{t,y}(X_t) \right]
\end{aligned}$$

$$+\sum_{t=1}^{T}\frac{(1-\alpha_t)^2}{3!{\alpha_t}^{3/2}}\mathbb{E}_{X_t\sim Q_{t|y}}\Big[3\sum_{i=1}^{d}\partial^3_{iii}\log q_{t-1|y}(m_{t,y}(X_t))\Delta_{t,y}(X_t)^i + \sum_{\substack{i,j=1\\ i\neq j}}^{d}\partial^3_{iij}\log q_{t-1|y}(m_{t,y}(X_t))\Delta_{t,y}(X_t)^j\Big]$$
$$+\max_{t\geq 1}\sqrt{\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}(X_t)\right\|^2}(\log T)\varepsilon.$$

From (Liang et al., 2024, Theorem 2) (and by assumption $N \leq d$), if the α_t satisfies Definition 1,

$$\mathcal{W}_{\text{oracle}} \lesssim \frac{d^2(\log T)^2}{T} + (\log T)\varepsilon^2.$$

Also, from Proposition 1, under the assumption on α_t, $\mathcal{W}_{\text{bias}}$ can be upper-bounded as

$$\mathcal{W}_{\text{bias}} \lesssim d + \sum_{n=1}^{N}\pi_n\left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2.$$

Among the terms in $\mathcal{W}_{\text{vanish}}$, the last estimation error term can be upper-bounded using Proposition 1 as

$$\max_{t\geq 1}\sqrt{\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}(X_t)\right\|^2}(\log T)\varepsilon \lesssim \sqrt{d + \sum_{n=1}^{N}\pi_n\left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2}(\log T)\varepsilon.$$

It remains to analyze the rest of the terms in $\mathcal{W}_{\text{vanish}}$. In the following we fix $t \geq 1$. We remind readers of the notations in (38). For the first term in $\mathcal{W}_{\text{vanish}}$, we first provide the following useful calculations. Note that by exchanging the order of expectation, for any function fn we have

$$\mathbb{E}_{X_t\sim Q_{t|y}}\left[\sum_{n=1}^{N}\pi_n\frac{q_{t,n|y}(X_t)}{q_{t|y}(X_t)}\mathrm{fn}(X_t,n)\right] = \sum_{n=1}^{N}\pi_n\mathbb{E}_{X_t\sim Q_{t,n|y}}\mathrm{fn}(X_t,n).$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}}\sum_{n=1}^{N}\pi_n\frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}\left|(X_t-\sqrt{\bar\alpha_t}\mu_{0,n})^\intercal X_t\right| \\
&= \sum_{n=1}^{N}\pi_n\mathbb{E}_{X_t\sim Q_{t,n|y}}\left|(X_t-\sqrt{\bar\alpha_t}\mu_{0,n})^\intercal X_t\right| \\
&\leq \sum_{n=1}^{N}\pi_n\mathbb{E}_{X_t\sim Q_{t,n|y}}\left\|X_t-\sqrt{\bar\alpha_t}\mu_{0,n}\right\|^2 + \sqrt{\bar\alpha_t}\sum_{n=1}^{N}\pi_n\left\|\mu_{0,n}\right\|\sqrt{\mathbb{E}_{X_t\sim Q_{t,n|y}}\left\|X_t-\sqrt{\bar\alpha_t}\mu_{0,n}\right\|^2} \\
&\overset{(i)}{=} \mathrm{Tr}(\Sigma_{t|y}) + \bar\alpha_t\sum_{n=1}^{N}\pi_n\left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2 \\
&\qquad + \sqrt{\bar\alpha_t}\sum_{n=1}^{N}\pi_n\left\|\mu_{0,n}\right\|\sqrt{\mathrm{Tr}(\Sigma_{t|y}) + \bar\alpha_t\sum_{n=1}^{N}\pi_n\left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2} \\
&\lesssim d + \sum_{n=1}^{N}\pi_n\left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2
\end{aligned} \tag{42}$$

where (i) follows from (58).

Also, note that $(I_d - H^\dagger H)\Sigma^{-r}_{t-1|y}(H^\dagger H) = 0$ using the following simple induction argument. For the base case, we have $(I_d - H^\dagger H)\Sigma^{-1}_{t-1|y}(H^\dagger H) = 0$ from Lemma 9. Then, suppose

$(I_d - H^\dagger H)\Sigma_{t-1|y}^{-(r-1)}(H^\dagger H) = 0$, we have $(I_d - H^\dagger H)\Sigma_{t-1|y}^{-r}(H^\dagger H) = (I_d - H^\dagger H)\Sigma_{t-1|y}^{-(r-1)}(I_d - H^\dagger H + H^\dagger H)\Sigma_{t-1|y}^{-1}(H^\dagger H) = (I_d - H^\dagger H)\Sigma_{t-1|y}^{-(r-1)}(I_d - H^\dagger H)\Sigma_{t-1|y}^{-1}(H^\dagger H) + (I_d - H^\dagger H)\Sigma_{t-1|y}^{-(r-1)}(H^\dagger H)\Sigma_{t-1|y}^{-1}(H^\dagger H) = 0$. Thus, for all $r \geq 1$ and any fixed vector v, with the definition of $\Delta_{t,y}$ in (40) and (41),

$$
\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t|y}} \left| (\Sigma_{t-1|y}^{-r} v)^\intercal \Delta_{t,y} \right| \\
&\leq \|v\| \, \mathbb{E}_{X_t \sim Q_{t|y}} \left\| \Sigma_{t-1|y}^{-r} \Delta_{t,y} \right\| = \|v\| \, \mathbb{E}_{X_t \sim Q_{t|y}} \left\| A_{t-1}^r \Delta_{t,y} \right\| \\
&\leq 4 \|v\| \left\| (\bar{\alpha}_{t-1}[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_{t-1}) I_{d-p})^{-r} \right\| \max_{n \in [N]} \left\| [\Sigma_t^{-1}]_{\bar{y}:} \mu_{0,n} \right\| \\
&\quad + 2 \|v\| \left\| (\bar{\alpha}_{t-1}[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_{t-1}) I_{d-p})^{-r} \right\| \left\| (\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} [\Sigma_0]_{\bar{y}y} [\Sigma_t^{-1}]_{y:} \right\| \times \\
&\qquad \sqrt{\mathbb{E}_{X_t \sim Q_{t|y}} \sum_{n=1}^N \pi_n \frac{q_{t,n|y}(X_t)}{q_{t|y}(X_t)} \left\| X_t - \sqrt{\bar{\alpha}_t} \mu_{0,n} \right\|^2} \\
&\lesssim d + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2
\end{aligned} \tag{43}
$$

where the last line follows from (58). Similarly,

$$
\mathbb{E}_{X_t \sim Q_{t|y}} \left| (\Sigma_{t|y}^{-r} v)^\intercal \Delta_{t,y} \right| \lesssim d + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2 .
$$

Also, for all $r \geq 1$,

$$
\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t|y}} \left| (\Sigma_{t-1|y}^{-r} X_t)^\intercal \Delta_{t,y} \right| = \mathbb{E}_{X_t \sim Q_{t|y}} \left| X_t^\intercal \Sigma_{t-1|y}^{-r} (I_d - H^\dagger H) \Delta_{t,y} \right| \\
&= \mathbb{E}_{X_t \sim Q_{t|y}} \left| X_t^\intercal A_{t-1}^r \Delta_{t,y} \right| \\
&\leq 4 \left\| (\bar{\alpha}_{t-1}[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_{t-1}) I_{d-p})^{-r} \right\| \max_{n \in [N]} \left\| [\Sigma_t^{-1}]_{\bar{y}:} \mu_{0,n} \right\| \mathbb{E}_{X_t \sim Q_{t|y}} \|X_t\| \\
&\quad + 2 \left\| (\bar{\alpha}_{t-1}[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_{t-1}) I_{d-p})^{-r} \right\| \left\| (\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} [\Sigma_0]_{\bar{y}y} [\Sigma_t^{-1}]_{y:} \right\| \times \\
&\qquad \mathbb{E}_{X_t \sim Q_{t|y}} \sum_{n=1}^N \pi_n \frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)} \left| (X_t - \sqrt{\bar{\alpha}_t} \mu_{0,n})^\intercal X_t \right| \\
&\lesssim d + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2
\end{aligned} \tag{44}
$$

where the last line follows from (42) and the fact that, from (58),

$$
\begin{aligned}
\sqrt{d} \cdot \mathbb{E}_{X_t \sim Q_{t|y}} \|X_t\| &\leq \sqrt{d} \sum_{n=1}^N \pi_n \sqrt{2 \mathbb{E}_{X_t \sim Q_{t,n|y}} \|X_t - \mu_{t,n}\|^2 + 2 \|\mu_{t,n}\|^2} \\
&\lesssim \sqrt{d} \sum_{n=1}^N \pi_n \sqrt{\mathrm{Tr}(\Sigma_{t|y}) + \bar{\alpha}_t \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2 + d} \\
&\lesssim d + \sqrt{d} \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\| \\
&\lesssim d + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2 .
\end{aligned}
$$

Similarly, we also have $\mathbb{E}_{X_t \sim Q_{t|y}} \left| (\Sigma_{t|y}^{-r} X_t)^\intercal \Delta_{t,y} \right| \lesssim d + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2$.

Also, for all $r \geq 1$, using the expression of $\nabla \log q_{t|y}$ in (39), and noting that by definition $(I_d - H^\dagger H)(x_t - \mu_{t,n|y}) = (I_d - H^\dagger H)(x_t - \mu_{t,n})$, we have

$$
\mathbb{E}_{X_t \sim Q_{t|y}} \left| (\Sigma_{t-1|y}^{-r} \nabla \log q_{t|y}(X_t))^\intercal \Delta_{t,y} \right|
$$

$$
\begin{aligned}
&= \mathbb{E}_{X_t\sim Q_{t|y}}\left|\left(A_{t-1}^r\sum_{n=1}^N \pi_n \frac{q_{t,n|y}(X_t)}{q_{t|y}(X_t)} A_t(I_d - H^\dagger H)(X_t - \mu_{t,n|y})\right)^\intercal \Delta_{t,y}\right| \\
&\le 4\left\|(\bar\alpha_{t-1}[\Sigma_0]_{\bar y\bar y} + (1-\bar\alpha_{t-1})I_{d-p})^{-r}\right\| \left\|(\bar\alpha_t[\Sigma_0]_{\bar y\bar y} + (1-\bar\alpha_t)I_{d-p})^{-1}\right\| \times \\
&\quad \mathbb{E}_{X_t\sim Q_{t|y}}\left\|\sum_{\ell=1}^N \pi_\ell \frac{q_{t,\ell|y}(X_t)}{q_{t|y}(X_t)}[X_t - \sqrt{\bar\alpha_t}\mu_{0,\ell}]_{\bar y}\right\| \times \max_{n\in[N]}\left\|[\Sigma_t^{-1}]_{\bar y:}\mu_{0,n}\right\| \\
&\quad + 2\left\|(\bar\alpha_{t-1}[\Sigma_0]_{\bar y\bar y} + (1-\bar\alpha_{t-1})I_{d-p})^{-r}\right\| \left\|(\bar\alpha_t[\Sigma_0]_{\bar y\bar y} + (1-\bar\alpha_t)I_{d-p})^{-1}\right\|^2 \times \\
&\quad \mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ N,L\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left[[X_t - \sqrt{\bar\alpha_t}\mu_{0,L}]_{\bar y}^\intercal [\Sigma_0]_{\bar y y}[\Sigma_t^{-1}]_{y:}(X_t - \sqrt{\bar\alpha_t}\mu_{0,N})\right] \\
&\lesssim \sqrt{\sum_{n=1}^N \pi_n \mathbb{E}_{X_t\sim Q_{t,n|y}}\left\|X_t - \sqrt{\bar\alpha_t}\mu_{0,n}\right\|^2} \times \sqrt{d} \\
&\quad + \mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ N,L\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left\|X_t - \sqrt{\bar\alpha_t}\mu_{0,L}\right\|\left\|X_t - \sqrt{\bar\alpha_t}\mu_{0,N}\right\| \\
&\lesssim d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2
\end{aligned} \tag{45}
$$

where the last line follows because, from (58),

$$
\begin{aligned}
&\mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ N,L\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left[\left\|X_t - \sqrt{\bar\alpha_t}\mu_{0,L}\right\|\left\|X_t - \sqrt{\bar\alpha_t}\mu_{0,N}\right\|\right] \\
&\le \sqrt{\mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ L\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left\|X_t - \sqrt{\bar\alpha_t}\mu_{0,L}\right\|^2} \times \sqrt{\mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ N\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left\|X_t - \sqrt{\bar\alpha_t}\mu_{0,N}\right\|^2} \\
&= \mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ N\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left\|X_t - \sqrt{\bar\alpha_t}\mu_{0,N}\right\|^2 \\
&\lesssim d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2.
\end{aligned}
$$

Now, we start to analyze the first term of $\mathcal{W}_{\text{vanish}}$. Recall that $m_{t,y}(x_t) = \mathbb{E}_{X_{t-1}\sim Q_{t-1|t,y}}[X_{t-1}] = \frac{1}{\sqrt{\alpha_t}}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log q_{t|y}(x_t)$. Using the score expressions in (39), we can calculate that given x_t (and thus $m_{t,y} = m_{t,y}(x_t)$),

$$
\begin{aligned}
&\nabla\log q_{t-1|y}(m_{t,y}) - \sqrt{\alpha_t}\nabla\log q_{t|y}(x_t) \\
&= \sqrt{\alpha_t}\Sigma_{t|y}^{-1}x_t - \Sigma_{t-1|y}^{-1}m_{t,y} \\
&\quad - \sqrt{\alpha_t}\Sigma_{t|y}^{-1}\sum_{n=1}^N \pi_n \frac{q_{t,n|y}(x_t)}{q_{t|y}(x_t)}\mu_{t,n|y} + \Sigma_{t-1|y}^{-1}\sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})}\mu_{t-1,n|y} \\
&= \left(\Sigma_{t|y}^{-1} - \Sigma_{t-1|y}^{-1}\right)\sqrt{\alpha_t}x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\Sigma_{t-1|y}^{-1}(x_t + \nabla\log q_{t|y}(x_t)) \\
&\quad + (1-\alpha_t)\Sigma_{t-1|y}^{-1}\sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})}\mu_{t-1,n|y} \\
&\quad + \alpha_t \frac{\sum_{n,\ell=1}^N \pi_n\pi_\ell\left(q_{t-1,n|y}(m_{t,y})q_{t,\ell|y}(x_t)\Sigma_{t-1|y}^{-1} - q_{t-1,\ell|y}(m_{t,y})q_{t,n|y}(x_t)\Sigma_{t|y}^{-1}\right)\mu_{t-1,n|y}}{q_{t-1|y}(m_{t,y})q_{t|y}(x_t)}.
\end{aligned}
$$

Here, using similar analyses as in the proof of Lemma 5, we get

$$
\begin{aligned}
&(m_{t,y} - \mu_{t-1,n|y}) - (x_t - \mu_{t,n|y}) = \frac{1-\sqrt{\alpha_t}}{\sqrt{\alpha_t}}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log q_{t|y}(x_t) - (1-\sqrt{\alpha_t})\mu_{t-1,n|y} \\
&\Sigma_{t|y}^{-1} - \Sigma_{t-1|y}^{-1} = \frac{1-\alpha_t}{\alpha_t}\Sigma_{t-1,n}^{-1} + \frac{1-\alpha_t}{\alpha_t^2}\Sigma_{t-1,n}^{-2} + O((1-\alpha_t)^2)
\end{aligned}
$$

$$
\begin{aligned}
&q_{t-1,n|y}(m_{t,y})q_{t,\ell|y}(x_t)\Sigma_{t-1|y}^{-1} - q_{t-1,\ell|y}(m_{t,y})q_{t,n|y}(x_t)\Sigma_{t|y}^{-1} \\
&= q_{t-1,n|y}(m_{t,y})q_{t,\ell|y}(x_t)(\Sigma_{t-1|y}^{-1} - \Sigma_{t|y}^{-1}) \\
&\quad + (q_{t-1,n|y}(m_{t,y})q_{t,\ell|y}(x_t) - q_{t-1,\ell|y}(m_{t,y})q_{t,n|y}(x_t))\Sigma_{t|y}^{-1} \\
&= q_{t-1,n|y}(m_{t,y})q_{t,\ell|y}(x_t)(\Sigma_{t-1|y}^{-1} - \Sigma_{t|y}^{-1}) \\
&\quad + \Bigg(\frac{1}{2}((m_{t,y} - \mu_{t-1,\ell|y}) - (x_t - \mu_{t,\ell|y}))^\intercal \Sigma_{t-1|y}^{-1}(m_{t,y} - \mu_{t-1,\ell|y}) \\
&\qquad + \frac{1}{2}(x_t - \mu_{t,\ell|y})^\intercal(\Sigma_{t-1|y}^{-1} - \Sigma_{t|y}^{-1})(m_{t,y} - \mu_{t-1,\ell|y}) \\
&\qquad + \frac{1}{2}(x_t - \mu_{t,\ell|y})^\intercal \Sigma_{t|y}^{-1}((m_{t,y} - \mu_{t-1,\ell|y}) - (x_t - \mu_{t,\ell|y})) \\
&\qquad - \frac{1}{2}((m_{t,y} - \mu_{t-1,n|y}) - (x_t - \mu_{t,n|y}))^\intercal \Sigma_{t-1|y}^{-1}(m_{t,y} - \mu_{t-1,n|y}) \\
&\qquad - \frac{1}{2}(x_t - \mu_{t,n|y})^\intercal(\Sigma_{t-1|y}^{-1} - \Sigma_{t|y}^{-1})(m_{t,y} - \mu_{t-1,n|y}) \\
&\qquad - \frac{1}{2}(x_t - \mu_{t,n|y})^\intercal \Sigma_{t|y}^{-1}((m_{t,y} - \mu_{t-1,n|y}) - (x_t - \mu_{t,n|y}))\Bigg)\Sigma_{t|y}^{-1} \\
&\quad + O((1-\alpha_t)^2)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\left|\mathbb{E}_{X_t\sim Q_{t|y}}\left[(\nabla \log q_{t-1|y}(m_{t,y}) - \sqrt{\alpha_t}\nabla \log q_{t|y})^\intercal \Delta_{t,y}\right]\right| \\
&\le \mathbb{E}_{X_t\sim Q_{t|y}}\left[\left|X_t^\intercal(\Sigma_{t|y}^{-1} - \Sigma_{t-1|y}^{-1})\Delta_{t,y}\right|\right] \\
&\quad + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\mathbb{E}_{X_t\sim Q_{t|y}}\left[\left|(X_t + \nabla \log q_{t|y})^\intercal(\Sigma_{t-1|y}^{-1})\Delta_{t,y}\right|\right] \\
&\quad + (1-\alpha_t)\mathbb{E}_{X_t\sim Q_{t|y}}\left[\left|\left(\sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})}\mu_{t-1,n|y}\right)^\intercal \Sigma_{t-1|y}^{-1}\Delta_{t,y}\right|\right] \\
&\quad + \mathbb{E}_{X_t\sim Q_{t|y}}\left|\Delta_{t,y}^\intercal \sum_{n,\ell=1}^N \pi_n\pi_\ell \left(\frac{q_{t-1,n|y}(m_{t,y})q_{t,\ell|y}(X_t)}{q_{t-1|y}(m_{t,y})q_{t|y}(X_t)}\Sigma_{t-1|y}^{-1} - \frac{q_{t-1,\ell|y}(m_{t,y})q_{t,n|y}(X_t)}{q_{t-1|y}(m_{t,y})q_{t|y}(X_t)}\Sigma_{t|y}^{-1}\right)\mu_{t-1,n|y}\right|.
\end{aligned}
$$

Among the four terms above, the first term $\lesssim d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2$ from (44) (along with the similar result for $\Sigma_{t|y}^{-1}$), the second term $\lesssim d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2$ from (44) and (45), and both the third and the fourth term $\lesssim d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2$ from (43) (along with the similar result for $\Sigma_{t|y}^{-1}$). Thus,

$$
\left|\mathbb{E}_{X_t\sim Q_{t|y}}\left[(\nabla \log q_{t-1|y}(m_{t,y}) - \sqrt{\alpha_t}\nabla \log q_{t|y}(X_t))^\intercal \Delta_{t,y}\right]\right| \lesssim d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2,
$$

and the first term in $\mathcal{W}_{\text{vanish}}$ satisfies that

$$
\begin{aligned}
&\sum_{t=1}^T \frac{1-\alpha_t}{\sqrt{\alpha_t}}\mathbb{E}_{X_t\sim Q_{t|y}}\left[(\nabla \log q_{t-1|y}(m_{t,y}(X_t)) - \sqrt{\alpha_t}\nabla \log q_{t|y}(X_t))^\intercal \Delta_{t,y}(X_t)\right] \\
&\qquad\qquad \lesssim \left(d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^2\right)\frac{\log(1/\delta)^2(\log T)^2}{T}.
\end{aligned}
$$

For the second term in $\mathcal{W}_{\text{vanish}}$, we first provide the following useful calculation. Similar to (43), for all $r \ge 1$ and any fixed vector v,

$$
\mathbb{E}_{X_t\sim Q_{t|y}}\left|(\Sigma_{t|y}^{-r}v)^\intercal \Delta_{t,y}\right|^2
$$

$$
\begin{aligned}
&\leq \|v\|^2 \, \mathbb{E}_{X_t \sim Q_{t|y}} \left\| \Sigma_{t|y}^{-r} \Delta_{t,y} \right\|^2 = \|v\|^2 \, \mathbb{E}_{X_t \sim Q_{t|y}} \left\| A_t^r \Delta_{t,y} \right\|^2 \\
&\lesssim \|v\|^2 \left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-r} \right\|^2 \max_{n \in [N]} \left\| [\Sigma_t^{-1}]_{\bar{y}:} \mu_{0,n} \right\|^2 \\
&\quad + \|v\|^2 \left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-r} \right\|^2 \left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} [\Sigma_0]_{\bar{y}y} [\Sigma_t^{-1}]_{y:} \right\|^2 \times \\
&\qquad \mathbb{E}_{X_t \sim Q_{t|y}} \sum_{n=1}^N \pi_n \frac{q_{t,n|y}(X_t)}{q_{t|y}(X_t)} \left\| X_t - \sqrt{\bar{\alpha}_t} \mu_{0,n} \right\|^2 \\
&\lesssim d^2 + d \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2
\end{aligned} \tag{46}
$$

where the last line follows from (58).

Also, similar to (44), for all $r \geq 1$,

$$
\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t|y}} \left| (\Sigma_{t|y}^{-r} m_{t,y})^\intercal \Delta_{t,y} \right|^2 \\
&\lesssim \mathbb{E}_{X_t \sim Q_{t|y}} \left| X_t^\intercal \Sigma_{t|y}^{-r} \Delta_{t,y} \right|^2 + (1-\alpha_t) \mathbb{E}_{X_t \sim Q_{t|y}} \left| (\nabla \log q_{t|y}(X_t))^\intercal \Sigma_{t|y}^{-r} \Delta_{t,y} \right|^2 \\
&\overset{(ii)}{\lesssim} \mathbb{E}_{X_t \sim Q_{t|y}} \left| X_t^\intercal \Sigma_{t|y}^{-r} (I_d - H^\dagger H) \Delta_{t,y} \right|^2 = \mathbb{E}_{X_t \sim Q_{t|y}} \left| X_t^\intercal A_t^r \Delta_{t,y} \right|^2 \\
&\lesssim \left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-r} \right\|^2 \max_{n \in [N]} \left\| [\Sigma_t^{-1}]_{\bar{y}:} \mu_{0,n} \right\|^2 \mathbb{E}_{X_t \sim Q_{t|y}} \left\| X_t \right\|^2 \\
&\quad + \left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-r} \right\|^2 \left\| (\bar{\alpha}_t [\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t) I_{d-p})^{-1} [\Sigma_0]_{\bar{y}y} [\Sigma_t^{-1}]_{y:} \right\|^2 \times \\
&\qquad \mathbb{E}_{X_t \sim Q_{t|y}} \sum_{n=1}^N \pi_n \frac{q_{t,n|y}(X_t)}{q_{t|y}(X_t)} \left| (X_t - \sqrt{\bar{\alpha}_t} \mu_{0,n})^\intercal X_t \right|^2 \\
&\lesssim d^2 + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^4
\end{aligned} \tag{47}
$$

where (ii) follows from the fact that $\mathbb{E}_{X_t \sim Q_{t|y}} \left| (\nabla \log q_{t|y}(X_t))^\intercal \Sigma_{t|y}^{-r} \Delta_{t,y} \right|^2 \lesssim d^2$ (using a similar argument for deriving (45)), and the last line follows because

$$
\begin{aligned}
d \cdot \mathbb{E}_{X_t \sim Q_{t|y}} \left\| X_t \right\|^2 &\leq d \sum_{n=1}^N \pi_n \left(2 \mathbb{E}_{X_t \sim Q_{t,n|y}} \left\| X_t - \mu_{t,n} \right\|^2 + 2 \left\| \mu_{t,n} \right\|^2 \right) \\
&\overset{(iii)}{\lesssim} d \sum_{n=1}^N \pi_n \left(\mathrm{Tr}(\Sigma_{t|y}) + \bar{\alpha}_t \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2 + d \right) \\
&\lesssim d^2 + d \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2
\end{aligned}
$$

where (iii) follows from (58), and also

$$
\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t|y}} \sum_{n=1}^N \pi_n \frac{q_{t,n|y}(X_t)}{q_{t|y}(X_t)} \left| (X_t - \sqrt{\bar{\alpha}_t} \mu_{0,n})^\intercal X_t \right|^2 \\
&= \mathbb{E}_{\substack{X_t \sim Q_{t|y} \\ N \sim \Pi_{\cdot|t,y}(\cdot|X_t)}} \left((X_t - \sqrt{\bar{\alpha}_t} \mu_{0,N})^\intercal X_t \right)^2 \\
&\leq 2 \mathbb{E}_{\substack{X_t \sim Q_{t|y} \\ N \sim \Pi_{\cdot|t,y}(\cdot|X_t)}} \left\| X_t - \sqrt{\bar{\alpha}_t} \mu_{0,N} \right\|^4 + 2 \sqrt{\mathbb{E}_{\substack{X_t \sim Q_{t|y} \\ N \sim \Pi_{\cdot|t,y}(\cdot|X_t)}} \left\| X_t - \sqrt{\bar{\alpha}_t} \mu_{0,N} \right\|^4} \sqrt{\mathbb{E}_{N \sim \Pi} \left\| \mu_{0,N} \right\|^4} \\
&\lesssim d^2 + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^4
\end{aligned}
$$

where the last line above follows because for all $r \geq 1$ and each $n \in [N]$,

$$\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t,n|y}} \left\| X_t - \sqrt{\bar{\alpha}_t}\mu_{0,n} \right\|^r \\
&\lesssim \mathbb{E}_{X_t \sim Q_{t,n|y}} \left\| X_t - \sqrt{\bar{\alpha}_t}\mu_{0,n|y} \right\|^r + \left\| \sqrt{\bar{\alpha}_t}\mu_{0,n|y} - \sqrt{\bar{\alpha}_t}\mu_{0,n} \right\|^r \\
&\leq \left\| \Sigma_{t|y}^{\frac{1}{2}} \right\|^r \mathbb{E}_{X_t \sim Q_{t,n|y}} \left\| \Sigma_{t|y}^{-\frac{1}{2}} (X_t - \sqrt{\bar{\alpha}_t}\mu_{0,n|y}) \right\|^r + (\bar{\alpha}_t)^{r/2} \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^r \\
&\lesssim d^{r/2} + \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^r . \qquad (48)
\end{aligned}$$

Now we are ready to analyze the second term of $\mathcal{W}_{\text{vanish}}$. Note that

$$\begin{aligned}
&\nabla^2 \log q_{t-1|y}(m_{t,y}) \\
&= \sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})} \left(\Sigma_{t|y}^{-1}(m_{t,y} - \mu_{t,n|y})(m_{t,y} - \mu_{t,n|y})^\intercal \Sigma_{t|y}^{-1} \right) - \Sigma_{t|y}^{-1} \\
&\quad - \left(\sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})} \Sigma_{t|y}^{-1}(m_{t,y} - \mu_{t,n|y}) \right) \left(\sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})} \Sigma_{t|y}^{-1}(m_{t,y} - \mu_{t,n|y}) \right)^\intercal .
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t|y}} \left| \Delta_{t,y}^\intercal \nabla^2 \log q_{t-1|y}(m_{t,y}) \Delta_{t,y} \right| \\
&\leq 3\mathbb{E}_{X_t \sim Q_{t|y}} \left[\sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})} \left(\Delta_{t,y}^\intercal \Sigma_{t|y}^{-1}(m_{t,y} - \mu_{t,n|y}) \right)^2 \right] \\
&\quad + 3\mathbb{E}_{X_t \sim Q_{t|y}} \left| \Delta_{t,y}^\intercal \Sigma_{t|y}^{-1} \Delta_{t,y} \right| \\
&\quad + 3\mathbb{E}_{X_t \sim Q_{t|y}} \left(\sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})} \Delta_{t,y}^\intercal \Sigma_{t|y}^{-1}(m_{t,y} - \mu_{t,n|y}) \right)^2 \\
&\leq 3\mathbb{E}_{X_t \sim Q_{t|y}} \left| \Delta_{t,y}^\intercal \Sigma_{t|y}^{-1} \Delta_{t,y} \right| \\
&\quad + 6\mathbb{E}_{X_t \sim Q_{t|y}} \left[\sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})} \left(\Delta_{t,y}^\intercal \Sigma_{t|y}^{-1}(m_{t,y} - \mu_{t,n|y}) \right)^2 \right] .
\end{aligned}$$

To determine the rate of these two terms, we get

$$\mathbb{E}_{X_t \sim Q_{t|y}} \left| \Delta_{t,y}^\intercal \Sigma_{t|y}^{-1} \Delta_{t,y} \right| = \mathbb{E}_{X_t \sim Q_{t|y}} \left| \Delta_{t,y}^\intercal A_t \Delta_{t,y} \right| \lesssim d + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2 ,$$

and

$$\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t|y}} \sum_{n=1}^N \pi_n \frac{q_{t-1,n|y}(m_{t,y})}{q_{t-1|y}(m_{t,y})} \left| (m_{t,y} - \mu_{t,n|y})^\intercal \Sigma_{t|y}^{-1} \Delta_{t,y} \right|^2 \\
&\quad \leq \mathbb{E}_{X_t \sim Q_{t|y}} \max_{n \in [N]} \left| (m_{t,y} - \mu_{t,n|y})^\intercal \Sigma_{t|y}^{-1} \Delta_{t,y} \right|^2 \\
&\quad \lesssim \mathbb{E}_{X_t \sim Q_{t|y}} \left| m_{t,y}^\intercal \Sigma_{t|y}^{-1} \Delta_{t,y} \right|^2 + \mathbb{E}_{X_t \sim Q_{t|y}} \max_{n \in [N]} \left| \mu_{t,n|y}^\intercal \Sigma_{t|y}^{-1} \Delta_{t,y} \right|^2 \\
&\quad \lesssim N \left(d^2 + \sum_{n=1}^N \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^4 \right)
\end{aligned}$$

where the last line follows from (46) and (47). Thus, the second term of $\mathcal{W}_{\text{vanish}}$ satisfies that

$$\sum_{t=1}^T \frac{(1-\alpha_t)^2}{2\alpha_t} \mathbb{E}_{X_t \sim Q_{t|y}} \left| \Delta_{t,y}(X_t)^\intercal \nabla^2 \log q_{t-1|y}(m_{t,y}(X_t)) \Delta_{t,y}(X_t) \right|$$

$$\lesssim N\left(d^2+\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^4\right)\frac{c^2(\log T)^2}{T}.$$

For the third term of $\mathcal{W}_{\text{vanish}}$, we provide the following useful calculations. Denote $v^{\circ 3}$ as the element-wise (Hadamard) third power of a vector v. For each $n\in[N]$, we have

$$\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}}\left|(m_{t,y}-\mu_{t,n|y})^{\circ 3}(I_d-H^\dagger H)\Delta_{t,y}\right|=\mathbb{E}_{X_t\sim Q_{t|y}}\left|[\Delta_{t,y}]_{\bar y}^{\intercal}[m_{t,y}-\mu_{t,n}]_{\bar y}^{\circ 3}\right|\\
&\lesssim \mathbb{E}_{X_t\sim Q_{t|y}}\left|[\Delta_{t,y}]_{\bar y}^{\intercal}[X_t-\mu_{t,n}]_{\bar y}^{\circ 3}\right|+(1-\alpha_t)\sqrt{\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}\right\|^2}\sqrt{\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\nabla\log q_{t|y}(X_t)\right\|_6^6}\\
&\overset{(iv)}{\lesssim} \mathbb{E}_{X_t\sim Q_{t|y}}\left|[\Delta_{t,y}]_{\bar y}^{\intercal}[X_t-\mu_{t,n}]_{\bar y}^{\circ 3}\right|\\
&\lesssim \max_{\ell}\left\|[\Sigma_t^{-1}]_{\bar y:}\mu_{0,\ell}\right\|\sqrt{\mathbb{E}_{X_t\sim Q_{t|y}}\left\|X_t-\mu_{t,n}\right\|_6^6}+\mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ L\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left\|X_t-\mu_{t,L}\right\|_4^4\\
&\lesssim \sqrt{\mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ L\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left\|X_t-\mu_{t,L}\right\|^6}+\sqrt{\mathbb{E}_{L\sim\Pi}\left\|\mu_{t,n}-\mu_{t,L}\right\|^6}+\mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ L\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left\|X_t-\mu_{t,L}\right\|^4\\
&\lesssim d^2+\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^4
\end{aligned}\tag{49}$$

where (iv) follows from Lemma 6 (using the α_t in (8)) and (Liang et al., 2024, Lemma 15), and the last line follows from (48). With a similar argument,

$$\begin{aligned}
\mathbb{E}_{X_t\sim Q_{t|y}}\left|(m_{t,y}-\mu_{t,n|y})(I_d-H^\dagger H)\Delta_{t,y}\right|&\lesssim \mathbb{E}_{\substack{X_t\sim Q_{t|y}\\ L\sim\Pi_{\cdot|t,y}(\cdot|X_t)}}\left\|X_t-\mu_{t,L}\right\|^2\\
&\lesssim d+\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^2.
\end{aligned}\tag{50}$$

Now, employing the notations from (Liang et al., 2024, Section G.1), we define

$$z_{t,n}(x):=\Sigma_{t|y}^{-1}(x-\mu_{t,n|y}),\quad \xi_t(x,i):=\max_n\left|z_{t,n}^i(x)\right|,\quad \bar\Sigma_t^{ij}:=\max_n\left|[\Sigma_{t|y}^{-1}]^{ij}\right|.$$

When $H=(I_p\quad 0)$, we note that

$$\Sigma_{t|y}^{-1}=\begin{pmatrix}(1-\bar\alpha_t+\bar\alpha_t\sigma_y^2)^{-1}I_p & 0\\ 0 & (\bar\alpha_t[\Sigma_0]_{\bar y\bar y}+(1-\bar\alpha_t)I_{d-p})^{-1}\end{pmatrix}.$$

Thus, we have $\bar\Sigma_t^{ij}\equiv 0$ whenever $(i,j)\in[1,p]\times[p+1,d]$ or $(i,j)\in[p+1,d]\times[1,p]$ and $\max_{i,j\in[p+1,d]}\bar\Sigma_t^{ij}=O(1)$. Since $\sigma_y^2>0$, we also have $\left\|\Sigma_{t|y}^{-1}\right\|\lesssim 1$.

From (Liang et al., 2024, Section G.1.2), an upper bound for third-order partial derivatives is

$$\left|\partial_{ijk}^3\log q_{t|y}(x)\right|\le 6\xi_t(x,i)\xi_t(x,j)\xi_t(x,k)+2\bar\Sigma_t^{ij}\xi_t(x,k)+2\bar\Sigma_t^{ik}\xi_t(x,j)+2\bar\Sigma_t^{jk}\xi_t(x,i).$$

We also remind readers that $\Delta_{t,y}$ is supported on $\mathrm{range}(I_d-H^\dagger H)$, namely that $[\Delta_{t,y}]_y\equiv 0$.

Now, the third term of $\mathcal{W}_{\text{vanish}}$ can be upper-bounded as

$$\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i=1}^{d}\partial_{iii}^3\log q_{t-1|y}(m_{t,y}(X_t))\Delta_{t,y}(X_t)^i\right|\\
&=\mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i=p+1}^{d}\partial_{iii}^3\log q_{t-1|y}(m_{t,y}(X_t))\Delta_{t,y}(X_t)^i\right|\\
&\lesssim \mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i=p+1}^{d}\xi_t(m_{t,y}(X_t),i)^3\Delta_{t,y}(X_t)^i\right|+\mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i=p+1}^{d}\bar\Sigma_t^{ii}\xi_t(m_{t,y}(X_t),i)\Delta_{t,y}(X_t)^i\right|
\end{aligned}$$

$$\lesssim \sum_{n=1}^{N} \left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}\right\|^3 \mathbb{E}_{X_t\sim Q_{t|y}} \left|\sum_{i=p+1}^{d} (m_{t,y}(X_t)^i - \mu^i_{t,n|y})^3 \Delta_{t,y}(X_t)^i\right|$$
$$+ \sum_{n=1}^{N} \left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}\right\| \mathbb{E}_{X_t\sim Q_{t|y}} \left|\sum_{i=p+1}^{d} (m_{t,y}(X_t)^i - \mu^i_{t,n|y}) \Delta_{t,y}(X_t)^i\right|$$
$$\lesssim N\left(d^2 + \sum_{n=1}^{N} \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^4\right).$$

Here the last line follows from (49) and (50).

We provide the following useful calculations to upper-bound the fourth term of $\mathcal{W}_{\text{vanish}}$. First, for all $r \geq 1$ and any fixed vector v,

$$\begin{aligned}
\mathbb{E}_{X_t\sim Q_{t|y}} \|m_{t,y} - v\|^r &= \mathbb{E}_{X_t\sim Q_{t|y}} \left\|\frac{1}{\sqrt{\alpha_t}} X_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}} \nabla \log q_{t|y}(X_t) - v\right\|^r \\
&\lesssim \mathbb{E}_{X_t\sim Q_{t|y}} \|X_t - v\|^r + (1-\alpha_t)\mathbb{E}_{X_t\sim Q_{t|y}} \left\|\nabla \log q_{t|y}(X_t)\right\|^r \\
&\overset{(v)}{\lesssim} \mathbb{E}_{\substack{X_t\sim Q_{t|y} \\ L\sim \Pi_{\cdot|t,y}(\cdot|X_t)}} \|X_t - \mu_{t,L}\|^r + \mathbb{E}_{L\sim\Pi} \|\mu_{t,L} - v\|^r \\
&\lesssim d^{r/2} + \sum_{n=1}^{N} \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^r
\end{aligned} \tag{51}$$

where (v) follows from Lemma 6 (using the α_t in (8)) and (Liang et al., 2024, Lemma 15), and the last line follows from (48). Now,

$$\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}} \sum_{i=1}^{d} \xi_t(m_{t,y}, i)^2 \left|\sum_{j=p+1}^{d} \xi_t(m_{t,y}, j)\Delta_{t,y}(X_t)^j\right| \\
&\lesssim \sum_{n,\ell=1}^{N} \mathbb{E}_{X_t\sim Q_{t|y}} \left\|m_{t,y} - \mu_{t,n|y}\right\|^2 \left|(m_{t,y} - \mu_{t,\ell|y})^\intercal \Delta_{t,y}(X_t)\right| \\
&\leq \sum_{n,\ell=1}^{N} \sqrt{\mathbb{E}_{X_t\sim Q_{t|y}} \left\|m_{t,y} - \mu_{t,n|y}\right\|^4} \left(\mathbb{E}_{X_t\sim Q_{t|y}} \left\|m_{t,y} - \mu_{t,\ell|y}\right\|^4 \mathbb{E}_{X_t\sim Q_{t|y}} \|\Delta_{t,y}(X_t)\|^4\right)^{1/4} \\
&\lesssim N^2\left(d^2 + \sum_{n=1}^{N} \pi_n \left\|H^\dagger y - H^\dagger H\mu_{0,n}\right\|^4\right)
\end{aligned} \tag{52}$$

where the first inequality follows from $\left\|\Sigma^{-1}_{t|y}\right\| \lesssim 1$ and the last line follows from (51) and (59). Also,

$$\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}} \left|\sum_{i,j=p+1}^{d} \bar{\Sigma}^{ii}_t \xi_t(m_{t,y}, j)\Delta_{t,y}(X_t)^j\right| \\
&= \left|\sum_{i=p+1}^{d} \bar{\Sigma}^{ii}_t\right| \cdot \mathbb{E}_{X_t\sim Q_{t|y}} \left|\sum_{j=p+1}^{d} \xi_t(m_{t,y}, j)\Delta_{t,y}(X_t)^j\right| \\
&\lesssim \left|\sum_{i=p+1}^{d} \bar{\Sigma}^{ii}_t\right| \cdot \sum_{n=1}^{N} \mathbb{E}_{X_t\sim Q_{t|y}} \left|(m_{t,y} - \mu_{t,n})^\intercal \Delta_{t,y}(X_t)\right| \\
&\lesssim \left|\sum_{i=p+1}^{d} \bar{\Sigma}^{ii}_t\right| \cdot \sum_{n=1}^{N} \sqrt{\mathbb{E}_{X_t\sim Q_{t|y}} \|m_{t,y} - \mu_{t,n}\|^2} \sqrt{\mathbb{E}_{X_t\sim Q_{t|y}} \|\Delta_{t,y}(X_t)\|^2}
\end{aligned}$$

$$\lesssim Nd\left(d+\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^2\right) \tag{53}$$

where the last line follows from Proposition 1 and (51). Also,

$$\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i,j=p+1}^{d}\bar{\Sigma}_t^{ij}\xi_t(m_{t,y},i)\Delta_{t,y}(X_t)^j\right| \\
&\lesssim \mathbb{E}_{X_t\sim Q_{t|y}}\left(\sum_{i=p+1}^{d}\xi_t(m_{t,y},i)\right)\left|\sum_{j=p+1}^{d}\Delta_{t,y}(X_t)^j\right| \\
&\leq \sqrt{d\cdot\mathbb{E}_{X_t\sim Q_{t|y}}\left(\sum_{i=p+1}^{d}\xi_t(m_{t,y},i)^2\right)}\sqrt{d\cdot\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}(X_t)\right\|^2} \\
&\lesssim \sqrt{d\cdot\sum_{n=1}^{N}\mathbb{E}_{X_t\sim Q_{t|y}}\left\|m_{t,y}-\mu_{t,n}\right\|^2}\sqrt{d\cdot\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}(X_t)\right\|^2} \\
&\lesssim \sqrt{N}d\left(d+\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^2\right)
\end{aligned} \tag{54}$$

where the last line follows from Proposition 1 and (51).

Now, the fourth term of $\mathcal{W}_{\text{vanish}}$ can be upper-bounded as

$$\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i,j=1}^{d}\partial_{iij}^3\log q_{t-1|y}(m_{t,y}(X_t))\Delta_{t,y}(X_t)^j\right| \\
&=\mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i=1}^{d}\sum_{j=p+1}^{d}\partial_{iij}^3\log q_{t-1|y}(m_{t,y}(X_t))\Delta_{t,y}(X_t)^j\right| \\
&\lesssim \mathbb{E}_{X_t\sim Q_{t|y}}\sum_{i=1}^{d}\xi_t(m_{t,y}(X_t),i)^2\left|\sum_{j=p+1}^{d}\xi_t(m_{t,y}(X_t),j)\Delta_{t,y}(X_t)^j\right| \\
&\quad+\mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i,j=p+1}^{d}\bar{\Sigma}_t^{ii}\xi_t(m_{t,y}(X_t),j)\Delta_{t,y}(X_t)^j\right| \\
&\quad+\mathbb{E}_{X_t\sim Q_{t|y}}\left|\sum_{i,j=p+1}^{d}\bar{\Sigma}_t^{ij}\xi_t(m_{t,y}(X_t),i)\Delta_{t,y}(X_t)^j\right|.
\end{aligned}$$

For the three terms above, the first term $\lesssim N^2\left(d^2+\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^4\right)$ from (52), the second term $\lesssim Nd\left(d+\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^2\right)$ from (53), and the last term $\lesssim\sqrt{N}d\left(d+\sum_{n=1}^{N}\pi_n\left\|H^\dagger y-H^\dagger H\mu_{0,n}\right\|^2\right)$ from (54).

Thus, overall, with the α_t in (8) (cf. Lemma 6), the third and fourth terms give us

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{(1-\alpha_t)^2}{3!{\alpha_t}^{3/2}}\mathbb{E}_{X_t\sim Q_{t|y}}\Bigg[3\sum_{i=1}^{d}\partial_{iii}^3\log q_{t-1|y}(m_{t,y}(X_t))\Delta_{t,y}(X_t)^i \\
&\qquad+\sum_{\substack{i,j=1\\ i\neq j}}^{d}\partial_{iij}^3\log q_{t-1|y}(m_{t,y}(X_t))\Delta_{t,y}(X_t)^j\Bigg]
\end{aligned}$$

$$\lesssim N^2 \left(d^2 + \sum_{n=1}^{N} \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^4 \right) \frac{(\log T)^2}{T}.$$

Therefore, combining all the above, since N is constant,

$$\begin{aligned}\mathrm{KL}(Q_{0|y} \| \widehat{P}_{0|y}) \lesssim & \left(d + \sum_{n=1}^{N} \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2 \right) \\ & + \left(d^2 + \sum_{n=1}^{N} \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^4 \right) \frac{(\log T)^2}{T} \\ & + \sqrt{d + \sum_{n=1}^{N} \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2} (\log T) \varepsilon.\end{aligned}$$

Remark 2. In case of general $\sigma_y \geq 0$, the same upper bound can be applied to $\left\| \Sigma_{t-1|y}^{-1} \right\|$ as detailed in Remark 1. Thus, with the α_t in (8), we similarly have

$$\mathcal{W}_{\text{oracle}} \lesssim \frac{d^2 (\log T)^2 \log(1/\delta)^2}{T} + (\log T) \varepsilon^2.$$

The rest of the proof is similar. Combining with Lemma 7, we would finally obtain

$$\begin{aligned}\mathrm{KL}(Q_{1|y} \| \widehat{P}_{1|y}) \lesssim & \left(d + \sum_{n=1}^{N} \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2 \right) \left(1 - \frac{2 \log(1/\delta) \log T}{T} \right) \\ & + \left(d^2 + \sum_{n=1}^{N} \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^4 \right) \frac{(\log T)^2 \log(1/\delta)^2}{T} + \sqrt{d + \sum_{n=1}^{N} \pi_n \left\| H^\dagger y - H^\dagger H \mu_{0,n} \right\|^2} (\log T) \varepsilon.\end{aligned}$$

Here $\mathrm{W}_2(Q_{1|y}, Q_{0|y})^2 \lesssim \delta d$.

J Auxiliary Lemmas and Proofs in Section 4

J.1 Proof of Proposition 2

Given $Q_0 = \mathcal{N}(\mu_0, \Sigma_0)$, from the conditional forward model in (5), we can calculate

$$\begin{aligned}Q_t &= \mathcal{N}(\sqrt{\bar{\alpha}_t} \mu_0, \bar{\alpha}_t \Sigma_0 + (1 - \bar{\alpha}_t) I_d) =: \mathcal{N}(\mu_t, \Sigma_t) \\ Q_{t|y} &= \mathcal{N}(\sqrt{\bar{\alpha}_t} (I_d - H^\dagger H) \mu_0 + \sqrt{\bar{\alpha}_t} H^\dagger y, \\ &\qquad \bar{\alpha}_t (I_d - H^\dagger H) \Sigma_0 (I_d - H^\dagger H) + \bar{\alpha}_t \sigma_y^2 H^\dagger (H^\dagger)^\intercal + (1 - \bar{\alpha}_t) I_d).\end{aligned}$$

Note that when $H = (I_p \quad 0)$ and $\sigma_y^2 > 0$, $q_{0|y}$ exists. Define

$$\begin{aligned}\mu_{t|y} &:= \sqrt{\bar{\alpha}_t} (I_d - H^\dagger H) \mu_0 + \sqrt{\bar{\alpha}_t} H^\dagger y \\ \Sigma_{t,sig} &:= \bar{\alpha}_t (I_d - H^\dagger H) \Sigma_0 (I_d - H^\dagger H) + (1 - \bar{\alpha}_t) I_d \\ \Sigma_{t|y} &:= \Sigma_{t,sig} + \bar{\alpha}_t \sigma_y^2 H^\dagger (H^\dagger)^\intercal. \end{aligned} \tag{55}$$

Here $\Sigma_{t,sig}$ is the signal variance at time t, and $\Sigma_{t|y}$ is the total variance of the signal and the measurement noise. Note that when $H = (I_p \quad 0)$, $[\Sigma_{t|y}^{-1}]_{\bar{y}\bar{y}} = [\Sigma_{t,sig}^{-1}]_{\bar{y}\bar{y}}$. We also calculate the respective scores of Q_t and $Q_{t|y}$:

$$\nabla \log q_t(x_t) = -\Sigma_t^{-1} (x_t - \mu_t), \quad \nabla \log q_{t|y}(x_t) = -\Sigma_{t|y}^{-1} (x_t - \mu_{t|y}).$$

Since $f_{t,y} = f_{t,y}^*$ (defined in (10)), from (32), the bias at each time is equal to

$$\begin{aligned}\Delta_{t,y} &= (I_d - H^\dagger H)(\nabla \log q_{t|y}(x_t) - \nabla \log q_t(x_t)) \\ &= (I_d - H^\dagger H) \left(\Sigma_t^{-1} (x_t - \sqrt{\bar{\alpha}_t} \mu_0) - \Sigma_{t|y}^{-1} (x_t - \sqrt{\bar{\alpha}_t} \mu_{0|y}) \right)\end{aligned}$$

$$= (I_d - H^\dagger H)\Sigma_t^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_0) \\ - (I_d - H^\dagger H)\Sigma_{t|y}^{-1}(x_t - \sqrt{\bar{\alpha}_t}(I_d - H^\dagger H)\mu_0 - \sqrt{\bar{\alpha}_t}H^\dagger y).$$

Now, define

$$V_t := (H^\dagger H)\Sigma_0(I_d - H^\dagger H) + (I_d - H^\dagger H)\Sigma_0(H^\dagger H) + (H^\dagger H)\Sigma_0(H^\dagger H)$$
$$A_t := (I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H)$$

Thus, we have $\Sigma_t = \Sigma_{t,sig} + \bar{\alpha}_t V_t$ and $\Sigma_{t|y} = \Sigma_{t,sig} + \bar{\alpha}_t\sigma_y^2 H^\dagger(H^\dagger)^\intercal$. By Woodbury matrix identity, for any two matrices A and B, their sum can be inversed as $(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}$. Thus, we get

$$\Sigma_t^{-1} = (\Sigma_{t,sig} + \bar{\alpha}_t V_t)^{-1} = \Sigma_{t,sig}^{-1} - \bar{\alpha}_t\Sigma_{t,sig}^{-1}V_t\Sigma_t^{-1}$$
$$\Sigma_{t|y}^{-1} = (\Sigma_{t,sig} + \bar{\alpha}_t\sigma_y^2 H^\dagger(H^\dagger)^\intercal)^{-1} = \Sigma_{t,sig}^{-1} - \bar{\alpha}_t\sigma_y^2\Sigma_{t,sig}^{-1}H^\dagger(H^\dagger)^\intercal\Sigma_{t|y}^{-1}$$
$$\overset{(i)}{=} \Sigma_{t,sig}^{-1} - \bar{\alpha}_t\sigma_y^2\Sigma_{t,sig}^{-1}(H^\dagger H)\Sigma_{t|y}^{-1}$$

where (i) holds under assumption $H = (I_p \quad 0)$. Thus,

$$\begin{aligned}
\Delta_{t,y} &= (I_d - H^\dagger H)\left(\Sigma_t^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_0) - \Sigma_{t|y}^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_{0|y})\right) \\
&= (I_d - H^\dagger H)\left(\Sigma_{t,sig}^{-1} - \bar{\alpha}_t\Sigma_{t,sig}^{-1}V_t\Sigma_t^{-1}\right)(x_t - \sqrt{\bar{\alpha}_t}\mu_0) \\
&\quad - (I_d - H^\dagger H)\left(\Sigma_{t,sig}^{-1} - \bar{\alpha}_t\sigma_y^2\Sigma_{t,sig}^{-1}(H^\dagger H)\Sigma_{t|y}^{-1}\right)(x_t - \sqrt{\bar{\alpha}_t}(I_d - H^\dagger H)\mu_0 - \sqrt{\bar{\alpha}_t}H^\dagger y) \\
&\overset{(ii)}{=} (I_d - H^\dagger H)\left(\Sigma_{t,sig}^{-1} - \bar{\alpha}_t\Sigma_{t,sig}^{-1}V_t\Sigma_t^{-1}\right)(x_t - \sqrt{\bar{\alpha}_t}\mu_0) \\
&\quad - (I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(x_t - \sqrt{\bar{\alpha}_t}(I_d - H^\dagger H)\mu_0 - \sqrt{\bar{\alpha}_t}H^\dagger y) \\
&= -\bar{\alpha}_t(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}V_t\Sigma_t^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_0) \\
&\quad + (I_d - H^\dagger H)\Sigma_{t,sig}^{-1}\Big((I_d - H^\dagger H)(x_t - \sqrt{\bar{\alpha}_t}\mu_0) + H^\dagger H(x_t - \sqrt{\bar{\alpha}_t}\mu_0)\Big) \\
&\quad - (I_d - H^\dagger H)\Sigma_{t,sig}^{-1}\Big((I_d - H^\dagger H)(x_t - \sqrt{\bar{\alpha}_t}\mu_0) + ((H^\dagger H)x_t - \sqrt{\bar{\alpha}_t}H^\dagger y)\Big) \\
&\overset{(iii)}{=} -\bar{\alpha}_t(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}V_t\Sigma_t^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_0) \\
&\overset{(iv)}{=} -\bar{\alpha}_t(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H)\Sigma_0(H^\dagger H)\Sigma_t^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_0) \\
&= -\bar{\alpha}_t A_t\Sigma_0(H^\dagger H)\Sigma_t^{-1}(x_t - \sqrt{\bar{\alpha}_t}\mu_0)
\end{aligned} \tag{56}$$

where (ii)–(iv) hold because $(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(H^\dagger H) = 0$ by Lemma 9.

Now, since $H^\dagger H = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}$, we can re-express A_t and $\Delta_{t,y}$ as follows.

$$\begin{aligned}
A_t &= (I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & I_{d-p} \end{pmatrix}\begin{pmatrix} (1-\bar{\alpha}_t)I_p & 0 \\ 0 & \bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p} \end{pmatrix}^{-1}\begin{pmatrix} 0 & 0 \\ 0 & I_{d-p} \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 \\ 0 & (\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1} \end{pmatrix} \\
\Delta_{t,y} &= \begin{pmatrix} 0 \\ -\bar{\alpha}_t(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}[\Sigma_0]_{\bar{y}y}[\Sigma_t^{-1}]_{y:}(x_t - \sqrt{\bar{\alpha}_t}\mu_0) \end{pmatrix},
\end{aligned}$$

and we have

$$\mathbb{E}_{X_t\sim Q_{t|y}}\|\Delta_{t,y}\|^2 = \bar{\alpha}_t^2\mathbb{E}_{X_t\sim Q_{t|y}}\left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}[\Sigma_0]_{\bar{y}y}[\Sigma_t^{-1}]_{y:}(X_t - \sqrt{\bar{\alpha}_t}\mu_0)\right\|^2$$
$$= \bar{\alpha}_t^2\mathbb{E}_{X_t\sim Q_{t|y}}\mathrm{Tr}\Bigg((\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}[\Sigma_0]_{\bar{y}y}[\Sigma_t^{-1}]_{y:}(X_t - \sqrt{\bar{\alpha}_t}\mu_0)(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal$$

$$[\Sigma_t^{-1}]_{:y}[\Sigma_0]_{y\bar{y}}(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}\Big)$$

$$\overset{(v)}{\leq} \bar{\alpha}_t^2 \left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}\right\|^2 \left\|[\Sigma_t^{-1}]_{y:}\right\|^2 \left\|[\Sigma_0]_{y\bar{y}}[\Sigma_0]_{\bar{y}y}\right\| \mathbb{E}_{X_t\sim Q_{t|y}} \left\|X_t - \sqrt{\bar{\alpha}_t}\mu_0\right\|^2 \tag{57}$$

where (v) follows from the fact that $|\mathrm{Tr}(UV)| \leq \|U\| \mathrm{Tr}(V)$ if V is positive semi-definite. To analyze each norm above, denote $\lambda_1 \geq \cdots \geq \lambda_d > 0$ to be the eigenvalues of Σ_0, and note that $\|[\Sigma_0]_{\bar{y}y}\| \leq \|\Sigma_0\| = \lambda_1$. The largest eigenvalue of Σ_t^{-1} is $(\bar{\alpha}_t\lambda_d + (1-\bar{\alpha}_t))^{-1}$, and note that $\left\|[\Sigma_t^{-1}]_{y:}\right\| \leq \left\|\Sigma_t^{-1}\right\|$. Also, since $[\Sigma_0]_{\bar{y}\bar{y}}$ is positive semi-definite, denote $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_{d-p} > 0$ to be its eigenvalues, and thus

$$\left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}\right\| \leq \frac{1}{\bar{\alpha}_t\tilde{\lambda}_{d-p} + (1-\bar{\alpha}_t)} \leq \frac{1}{\min\{\tilde{\lambda}_{d-p}, 1\}} < \infty.$$

Also, since

$$\begin{aligned}
&\mathbb{E}_{X_t\sim Q_{t|y}}(X_t - \sqrt{\bar{\alpha}_t}\mu_0)(X_t - \sqrt{\bar{\alpha}_t}\mu_0)^\intercal \\
&= \mathbb{E}_{X_t\sim Q_{t|y}}(X_t - \mu_{t|y} + \sqrt{\bar{\alpha}_t}(H^\dagger y - H^\dagger H\mu_0))(X_t - \mu_{t|y} + \sqrt{\bar{\alpha}_t}(H^\dagger y - H^\dagger H\mu_0))^\intercal \\
&= \Sigma_{t|y} + \bar{\alpha}_t(H^\dagger y - H^\dagger H\mu_0)(H^\dagger y - H^\dagger H\mu_0)^\intercal,
\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{E}_{X_t\sim Q_{t|y}} \left\|X_t - \sqrt{\bar{\alpha}_t}\mu_0\right\|^2 &= \mathrm{Tr}(\Sigma_{t|y}) + \bar{\alpha}_t \left\|H^\dagger y - H^\dagger H\mu_0\right\|^2 \\
&\overset{(vi)}{\leq} \bar{\alpha}_t(\lambda_1 + \sigma_y^2) + (1-\bar{\alpha}_t) + \bar{\alpha}_t \left\|H^\dagger y - H^\dagger H\mu_0\right\|^2 \\
&\leq \max\left\{\left\|H^\dagger y - H^\dagger H\mu_0\right\|^2 + d(\lambda_1 + \sigma_y^2), d\right\}
\end{aligned} \tag{58}$$

where (vi) is because $\Sigma_{t|y}$ is positive definite with

$$\begin{aligned}
\left\|\Sigma_{t|y}\right\| &= \left\|\bar{\alpha}_t(I_d - H^\dagger H)\Sigma_0(I_d - H^\dagger H) + \bar{\alpha}_t\sigma_y^2 H^\dagger(H^\dagger)^\intercal + (1-\bar{\alpha}_t)I_d\right\| \\
&\leq \bar{\alpha}_t \|\Sigma_0\| + \bar{\alpha}_t\sigma_y^2 + (1-\bar{\alpha}_t) \\
&= \bar{\alpha}_t(\lambda_1 + \sigma_y^2) + (1-\bar{\alpha}_t).
\end{aligned}$$

Therefore, for the second moment,

$$\begin{aligned}
\mathbb{E}_{X_t\sim Q_{t|y}} \|\Delta_{t,y}\|^2 &\leq \bar{\alpha}_t^2 \frac{\max\left\{\left\|H^\dagger y - H^\dagger H\mu_0\right\|^2 + d(\lambda_1 + \sigma_y^2), d\right\}}{\min\{\lambda_d, 1\}^2 \min\{\tilde{\lambda}_{d-p}, 1\}^2} \left\|[\Sigma_0]_{y\bar{y}}[\Sigma_0]_{\bar{y}y}\right\| \\
&\lesssim \bar{\alpha}_t^2 \cdot \max\left\{\left\|H^\dagger y - H^\dagger H\mu_0\right\|^2 + d(\lambda_1 + \sigma_y^2), d\right\}
\end{aligned}$$

since $\left\|[\Sigma_0]_{y\bar{y}}[\Sigma_0]_{\bar{y}y}\right\| \leq \|\Sigma_0\|^2 = \lambda_1^2$. Also, for general moments with $m \geq 2$,

$$\begin{aligned}
\|\Delta_{t,y}\|^m &\leq \bar{\alpha}_t^m \left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}[\Sigma_0]_{\bar{y}y}[\Sigma_t^{-1}]_{y:}(x_t - \sqrt{\bar{\alpha}_t}\mu_0)\right\|^m \\
&\leq \bar{\alpha}_t^m \left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}\right\|^m \|\Sigma_0\|^m \left\|\Sigma_t^{-1}\right\|^m \left\|x_t - \sqrt{\bar{\alpha}_t}\mu_0\right\|^m \\
&\lesssim \bar{\alpha}_t^m \left\|x_t - \sqrt{\bar{\alpha}_t}\mu_0\right\|^m
\end{aligned}$$

and thus

$$\begin{aligned}
\mathbb{E}_{X_t\sim Q_{t|y}} \|\Delta_{t,y}\|^m &\lesssim \bar{\alpha}_t^m \left(\mathbb{E}_{X_t\sim Q_{t|y}} \left\|\Sigma_{t|y}^{-\frac{1}{2}}(X_t - \mu_{t|y})\right\|^m + \left\|\sqrt{\bar{\alpha}_t}(H^\dagger y - H^\dagger H\mu_0)\right\|^m\right) \\
&\leq \bar{\alpha}_t^m \left((m-1)!! \cdot d^{m/2-1} + \left\|\sqrt{\bar{\alpha}_t}(H^\dagger y - H^\dagger H\mu_0)\right\|^m\right) = O(\bar{\alpha}_t).
\end{aligned}$$

Therefore, Assumption 4 is satisfied if the α_t satisfies Definition 1. The proof is now complete.

J.2 PROOF OF LEMMA 8

We continue from the expression of $\Delta_{t,y}$ in (40) when Q_0 is Gaussian mixture. For $m \geq 2$, we have

$$
\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t|y}} \|\Delta_{t,y}\|^m \\
&\leq 2^{m-1}(\bar{\alpha}_t)^{m/2} \max_{n\in[N]} \left\|\Sigma_t^{-1}\mu_{0,n}\right\|^m \\
&\quad + 2^{m-1}(\bar{\alpha}_t)^m \mathbb{E}_{\substack{X_t\sim Q_{t|y} \\ N\sim \Pi_{\cdot|t,y}(\cdot|X_t)}} \left\|(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H)\Sigma_0(H^\dagger H)\Sigma_t^{-1}(X_t - \sqrt{\bar{\alpha}_t}\mu_{0,N})\right\|^m \\
&\overset{(ii)}{\lesssim} (\bar{\alpha}_t)^{m/2} d^{m/2} + (\bar{\alpha}_t)^m \mathbb{E}_{\substack{X_t\sim Q_{t|y} \\ N\sim \Pi_{\cdot|t,y}(\cdot|X_t)}} \left\|X_t - \sqrt{\bar{\alpha}_t}\mu_{0,N}\right\|^m \\
&= (\bar{\alpha}_t)^{m/2} d^{m/2} + (\bar{\alpha}_t)^m \mathbb{E}_{\substack{N\sim \Pi \\ X_t\sim Q_{t,N|y}}} \left\|X_t - \sqrt{\bar{\alpha}_t}\mu_{0,N}\right\|^m \\
&\overset{(iii)}{\lesssim} (\bar{\alpha}_t)^{m/2} d^{m/2} + (\bar{\alpha}_t)^m \left(d^{m/2} + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H \mu_{0,n}\right\|^m \right) \qquad (59) \\
&= O(\bar{\alpha}_t).
\end{aligned}
$$

Here (ii) follows because

$$
\begin{aligned}
\left\|(I_d - H^\dagger H)\Sigma_{t,sig}^{-1}(I_d - H^\dagger H)\Sigma_0(H^\dagger H)\Sigma_t^{-1}\right\| &= \left\|(\bar{\alpha}_t[\Sigma_0]_{\bar{y}\bar{y}} + (1-\bar{\alpha}_t)I_{d-p})^{-1}[\Sigma_0]_{\bar{y}y}[\Sigma_t^{-1}]_{y:}\right\| \\
&\leq \frac{\lambda_1}{\min\{\tilde{\lambda}_{d-p}, 1\}\min\{\lambda_d, 1\}} = O(1),
\end{aligned}
$$

and (iii) follows from (48). Therefore, this verifies Assumption 4 when the α_t satisfies Definition 1. The proof is now complete.

J.3 LEMMA 9 AND ITS PROOF

Lemma 9. *Given a positive semi-definite matrix Σ, $\sigma \geq 0$, and $\alpha \in (0,1)$,*

$$
\begin{aligned}
&(\alpha\sigma H^\dagger (H^\dagger)^\intercal + (1-\alpha)I_d)^{-1}(I_d - H^\dagger H) = \frac{1}{1-\alpha}(I_d - H^\dagger H) \\
&(I_d - H^\dagger H)(\alpha(I_d - H^\dagger H)\Sigma(I_d - H^\dagger H) + \alpha\sigma H^\dagger H + (1-\alpha)I_d)^{-1}(H^\dagger H) = 0.
\end{aligned}
$$

Proof. The key of the proof is the Woodbury matrix identity, which states that for any matrices $U \in \mathbb{R}^{d\times p}, V \in \mathbb{R}^{p\times d}$,

$$
(I_d + UV)^{-1} = I_d - U(I_p + VU)^{-1}V.
$$

For the first equality, we apply Woodbury with $U = \sqrt{\frac{\alpha\sigma}{1-\alpha}}H^\dagger$ and $V = \sqrt{\frac{\alpha\sigma}{1-\alpha}}(H^\dagger)^\intercal$ and we get

$$
(\alpha\sigma H^\dagger (H^\dagger)^\intercal + (1-\alpha)I_d)^{-1} = \frac{1}{1-\alpha}\left(I_d - \frac{\alpha\sigma}{1-\alpha}H^\dagger\left(I_p + \frac{\alpha\sigma}{1-\alpha}(H^\dagger)^\intercal H^\dagger\right)^{-1}(H^\dagger)^\intercal\right).
$$

Since $p \leq d$, the pseudo-inverse equals $H^\dagger = H^\intercal(HH^\intercal)^{-1}$, and by the orthogonal property we have

$$
(H^\dagger)^\intercal(I_d - H^\dagger H) = (HH^\intercal)^{-1}H(I_d - H^\dagger H) = 0.
$$

We have thus shown the first equality.

For the second equality, we first consider the case where $\sigma = 0$. Write $S = (1-\alpha)I_d + \alpha(I_d - H^\dagger H)\Sigma(I_d - H^\dagger H)$. Since Σ is positive semi-definite, there exists matrix L such that $\Sigma = LL^\intercal$. Thus,

$$
\begin{aligned}
S^{-1} &= \left((1-\alpha)I_d + \alpha(I_d - H^\dagger H)\Sigma(I_d - H^\dagger H)\right)^{-1} \\
&= \left((1-\alpha)I_d + \alpha\Big((I_d - H^\dagger H)L\Big)\Big((I_d - H^\dagger H)^\intercal L\Big)^\intercal\right)^{-1}
\end{aligned}
$$

$$= \frac{1}{1-\alpha} I_d - \frac{\alpha}{1-\alpha}((I_d - H^\dagger H)L)\left(I_d + \frac{\alpha}{1-\alpha} L^\intercal (I_d - H^\dagger H)L\right)^{-1} L^\intercal (I_d - H^\dagger H),$$

where in the last line we have applied Woodbury with $U = V^\intercal = \sqrt{\frac{\alpha}{1-\alpha}}(I_d - H^\dagger H)L$. The equality is achieved because

$$\begin{aligned}
&(I_d - H^\dagger H)S^{-1}(H^\dagger H) \\
&= \frac{1}{1-\alpha} \underbrace{(I_d - H^\dagger H)(H^\dagger H)}_{=0} \\
&\quad - \frac{\alpha}{1-\alpha}(I_d - H^\dagger H)L\left(I_d + \frac{\alpha}{1-\alpha} L^\intercal (I_d - H^\dagger H)L\right)^{-1} L^\intercal \underbrace{(I_d - H^\dagger H)(H^\dagger H)}_{=0} \\
&= 0.
\end{aligned}$$

□

When $\sigma > 0$, we can apply Woodbury identity to sum of matrices A and B and get $(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}$. Thus,

$$(S + \alpha\sigma H^\dagger H)^{-1} = S^{-1} - \alpha\sigma S^{-1}(H^\dagger H)(S + \alpha\sigma H^\dagger H)^{-1}$$

and

$$\begin{aligned}
&(I_d - H^\dagger H)(S + \alpha\sigma H^\dagger H)^{-1}(H^\dagger H) \\
&= \underbrace{(I_d - H^\dagger H)S^{-1}(H^\dagger H)}_{=0} \\
&\qquad - \alpha\sigma \underbrace{(I_d - H^\dagger H)S^{-1}(H^\dagger H)}_{=0}(S + \alpha\sigma H^\dagger H)^{-1}(H^\dagger H) \\
&= 0.
\end{aligned}$$

The proof is now complete.

J.4 Lemma 10 and its proof

Lemma 10. *With $1 - \alpha_t \equiv \frac{\log T}{T}$, $\forall t \geq 0$ (which satisfies Definition 1), given any $p > 0$,*

$$\sum_{t=2}^{T} (1-\alpha_t)\bar{\alpha}_t^p = \frac{1}{p}\left(1 - \frac{2pc\log T}{T}\right) + \tilde{O}\left(\frac{1}{T^2}\right).$$

Proof. Define the sum as s_T. Then,

$$\begin{aligned}
s_T &= \sum_{t=2}^{T} \frac{c\log T}{T}\left(1 - \frac{c\log T}{T}\right)^{pt} = \frac{c\log T}{T}\left(1 - \frac{c\log T}{T}\right)^{2p} \frac{1 - \left(1 - \frac{c\log T}{T}\right)^{p(T-1)}}{1 - \left(1 - \frac{c\log T}{T}\right)^{p}} \\
&= \frac{c\log T}{T}\left(1 - \frac{c\log T}{T}\right)^{2p} \frac{1}{1 - \left(1 - \frac{c\log T}{T}\right)^{p}}(1 + O(T^{-cp})) \\
&= \frac{1}{p}\left(1 - \frac{2pc\log T}{T}\right) + \tilde{O}(T^{-2}).
\end{aligned}$$

□