

1 Technical Appendices and Supplementary Material

2 Contents

3	A Limitations and future works	1
4	B Related works	2
5	B.1 Chance constrained programming	2
6	B.2 Optimization via sampling	2
7	B.3 Learning to optimize	2
8	B.4 Diffusion models for optimization	2
9	C Experimental details	3
10	C.1 Experimental settings	3
11	C.2 Effects of gradient guidance	4
12	C.3 Additional experimental results	5
13	C.3.1 Linear chance constrained problem	5
14	C.3.2 Computational cost	9
15	C.3.3 Variance schedule	9
16	C.3.4 Guidance term	9
17	C.3.5 VaR-constrained mean–variance portfolio selection problem	10
18	C.3.6 Robust waveform design	11
19	D Restricted problem	13
20	D.1 Connection with CCP	13
21	D.2 Special cases	17
22	E Technical appendices	19
23	E.1 Proof of Theorem 1	19
24	E.2 Proof of Corollary 1	20
25	E.3 Proof of Theorem 2	22
26	E.4 Proof of Theorem 3	23

27 A Limitations and future works

28 First, while empirical results demonstrate faster convergence with second-order guidance, theoretical
29 guarantees of this acceleration remain to be established. Second, while the U-Net architecture serves
30 as our baseline score estimator, it may not be optimal for all problem domains. Specialized network
31 architectures that better capture the geometric structure of constraints may be investigated. Third,
32 experimental results primarily focus on two specific types of problems, then further evaluation may
33 be required to assess the effectiveness on a broader range of function types.

34 B Related works

35 B.1 Chance constrained programming

36 CCP is a powerful modeling paradigm for optimization problems with uncertain constraints, with
37 applications across engineering, finance, and beyond. Two common solution approaches are Convex
38 Approximation (CA) and Sample Average Approximation (SAA). However, CA requires explicit
39 distributional information, and SAA can be computationally expensive. Thus, designing an efficient
40 framework for CCP under **unknown distributions** remains a pressing challenge.

41 B.2 Optimization via sampling

42 Traditional gradient-based methods often converge to local minima under nonconvex settings.
43 Sampling-based algorithms, particularly Langevin Dynamics, have demonstrated strong performance
44 in global optimization (Ma et al. [2019]). Compared to conventional optimizers, Sampling-based
45 algorithms can take full advantages of data priors and solve nonconvex problems more effectively.

46 B.3 Learning to optimize

47 In order to improve the efficiency of optimization algorithms, learning-based methods are studied
48 by Chen et al. [2022]. Learning-based methods aim to learn a parameterized or semi-parameterized
49 update rule of optimization without taking the form of any analytic update. Traditional learning-based
50 methods simply learn the mapping between the input and output of the optimization algorithms,
51 which may cause to fall into local minima. Consequently, generative sampling-based models have
52 attracted growing interest for optimization tasks.

53 B.4 Diffusion models for optimization

54 The rising prominence of diffusion models has spurred significant research interest in their underlying
55 mathematical foundations and theoretical properties, as well as strategies to optimize their perfor-
56 mance. At the same time, there are more and more researches on the application of diffusion model.
57 How to use diffusion model to solve optimization problems is gradually attracting people’s attention.
58 In Chung et al. [2022], an additional correction term inspired by the manifold constraint is added into
59 the reverse diffusion step to preserve the manifold constraint and data consistency, and used to solve
60 the inverse problem. In Krishnamoorthy et al. [2023], a conditional diffusion model is trained via
61 loss reweighting to map function values to corresponding points and applied for offline Black-Box
62 Optimization. In Guo et al. [2024], a kind of Look-Ahead Guidance (LAG) is introduced to preserve
63 the linear structure of data and then used for regularized optimization and global optimization. In Li
64 et al. [2024], a diffusion-based training-to-testing (T2T) framework is used to solve new instances in
65 combinatorial optimization while training on historical instances generated by existing algorithms.

66 Compared with related methods, our work is the first, to the best of our knowledge, to use diffusion
67 models to solve the general chance constrained problems. The key challenge here is the **lack of**
68 **direct training data** corresponding to the product distribution of the objective and constraints. We
69 address this through a dedicated data generation stage, followed by conditional training of the score.
70 In contrast, Guo et al. [2024] assumes access to a pre-trained unconditional diffusion model and
71 focuses on a restricted linear-Gaussian setting. Unlike classical convex approximation approaches for
72 CCP, our method does not require prior knowledge of the underlying distribution. Instead, we only
73 assume access to samples from it, which makes our approach applicable to broader and more realistic
74 settings.

75 More specifically, our approach introduces two main innovations:

- 76 • **Conditional Training and Applicability Beyond Linear-Gaussian Settings:** Unlike Guo et al.
77 [2024], which applies guidance to pre-trained unconditional diffusion models and assumes a linear
78 objective with Gaussian data, our framework involves a dedicated data generation process followed
79 by conditional score training. This enables us to address nonlinear and structurally complex
80 chance-constrained problems, where directly sampling from the feasible region is nontrivial.

- **A New Class of Guidance Derived from Product Distributions:** Most existing guided diffusion frameworks follow the general SDE form as follows:

$$d\mathbf{x}_t = [\mathbf{a}(\mathbf{x}_t, t) - b(t)^2(\mathbf{s}(\mathbf{x}_t, t) + \mathbf{G}_t)]dt + b(t)d\bar{\mathbf{B}}_t. \quad (1)$$

In our work, we derive two types of guidance terms directly from the product distribution formulation of the target density:

- a first-order guidance

$$\mathbf{G}_t^{(1)} = -\beta \nabla f(\mathbf{x}_t), \quad (2)$$

- a second-order guidance

$$\mathbf{G}_t^{(2)} = -\frac{1}{\sigma_{0|t}^2}[\mathbf{H}^{-1}[(\nabla_{\mathbf{x}_t}^2 f(\mathbf{x}_t)\mathbf{x}_t + \nabla f(\mathbf{x}_t)) - \frac{1}{\beta\sigma_{0|t}^2}\boldsymbol{\mu}_{0|t}] + \boldsymbol{\mu}_{0|t}], \quad (3)$$

where the terms are computed based on a learned surrogate for the chance constraint and the posterior mean $\boldsymbol{\mu}_{0|t}$.

In contrast, Guo et al. [2024] introduces a Look-Ahead Guidance term designed for linear objectives:

$$\mathbf{G}_t^{(3)} = -\beta(t)\nabla_{\mathbf{x}_t}(y - \mathbf{g}^\top \hat{\mathbb{E}}[\mathbf{x}_0|\mathbf{x}_t])^2, \quad (4)$$

where $\beta(t)$ and y are tuning parameters, \mathbf{g} is the gradient of the linear objective, and $\hat{\mathbb{E}}[\mathbf{x}_0|\mathbf{x}_t]$ is an approximation of the posterior mean $\boldsymbol{\mu}_{0|t}$ that can be calculated by the score network, i.e., $\hat{\mathbb{E}}[\mathbf{x}_0|\mathbf{x}_t] = \alpha^{-1}(t)(\mathbf{x}_t + h(t)\mathbf{s}_\theta(\mathbf{x}_t, t))$. This approach is effective when the data distribution is Gaussian and the objective is linear, but may degrade under nonlinear or non-Gaussian scenarios.

C Experimental details

C.1 Experimental settings

Our neural network architecture follows the backbone of a U-Net (Ronneberger et al. [2015]) and ResNet (He et al. [2016]). We use group normalization (Wu and He [2018]) to make the implementation simpler. All models use four feature map resolutions with convolutional residual blocks and self-attention blocks (Vaswani et al. [2017]) per resolution level. Diffusion time t and condition parameter ρ is specified by adding the Transformer sinusoidal position embedding into each residual block.

All models are trained with 4 A800 GPUs. The training durations are approximately 0.4 hours for the linear chance constrained problem and 2 hours for the robust waveform design task. The average sampling times are listed alongside the corresponding experimental results.

We set almost all our hyperparameters as default in (Ho et al. [2020], Guo et al. [2024]):

- We test the $\eta(t)$ schedule from a set of constant, linear, quadratic and cosine schedules. We set $T = 1000$ without a sweep and chose a linear schedule from $\eta(0) = 10^{-4}$ to $\eta(T) = 0.02$.
- We use Adam in our experimentation process and leave the hyperparameters to their standard values. We set the learning rate to 10^{-4} without any sweeping.
- We set the batch size to 64 for linear chance constrained problem and 128 for robust waveform design.

To generate the dataset, we utilize CVX (Grant et al. [2008]) to solve the restricted problem. For the linear chance constrained problem, we generate $N = 1000$ data samples, while $N = 10000$ samples for the robust waveform design task. During the sampling with guidance stage, we evaluate both first- and second-order gradient guidance by implementing a DDIM-based technique (Song et al. [2020a]) with a descaled time step $T' = 100$ to accelerate the sampling process.

Our code is available at <https://github.com/boyangzhang2000/GGDOpt>.

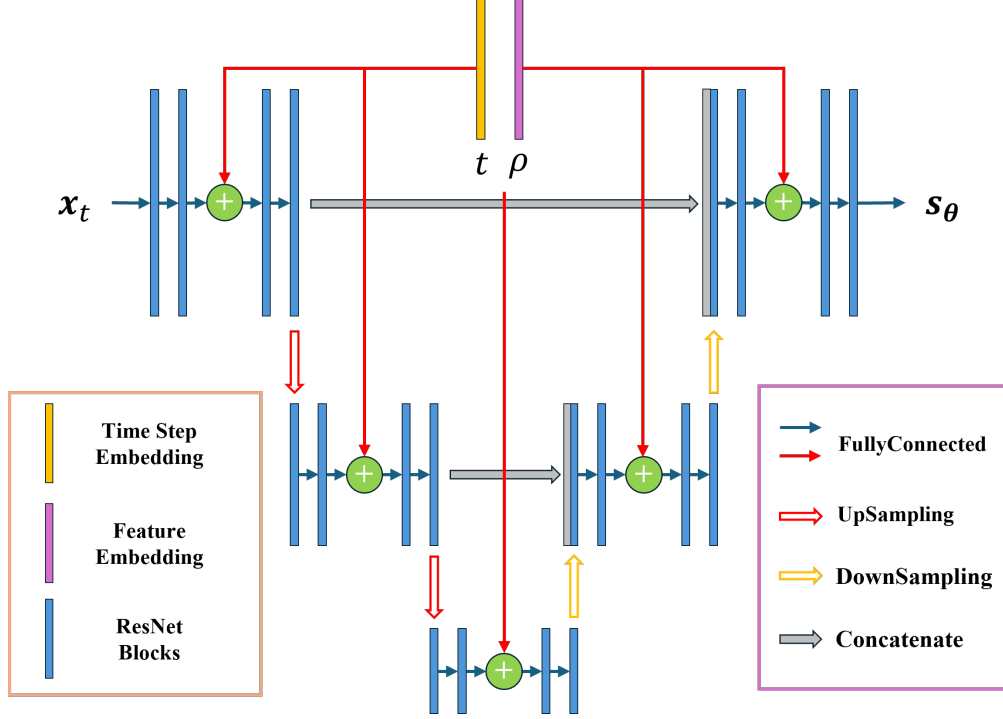
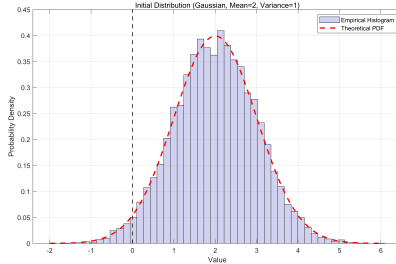


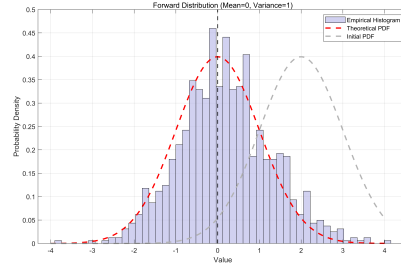
Figure 1: A sketch map for U-Net structure of GGDOpt

119 C.2 Effects of gradient guidance

120 First, we present an intuitive example illustrating how gradient guidance can steer the sampling
 121 trajectory toward the desired target. Specifically, we consider a one-dimensional sampling task where
 122 the initial distribution is $x_0 \sim \mathcal{N}(2, 1)$ and 1000 samples are drawn from it to serve as training data.
 123 We set the diffusion time step to $T = 1000$ and the resulting forward process of GGDOpt is shown to
 124 closely approximate the theoretical distribution $\mathcal{N}(0, 1)$ (see Figure 2).



(a) Initial distribution



(b) Forward distribution

Figure 2: The forward process of GGDOpt.

125 Next, we compare different sampling strategies: without guidance, first-order gradient guidance,
 126 and second-order gradient guidance. Theoretically, under Gaussian assumptions, first-order gradient
 127 guidance alters only the mean of the end distribution, whereas second-order gradient guidance affects
 128 both the mean and the variance. For each method, we generate 1000 samples and the corresponding
 129 sampling results are presented in Figure 3.

130 Experimental results demonstrate that, in the absence of guidance, the sampling process shifts the
 131 distribution from the prior $\mathcal{N}(0, 1)$ back to the initial distribution $\mathcal{N}(2, 1)$, as expected. When
 132 applying first-order gradient guidance with $\beta = 3$, the distribution transitions from the prior $\mathcal{N}(0, 1)$
 133 to the guided distribution $\mathcal{N}(5, 1)$, indicating a change in the mean while preserving the variance. In

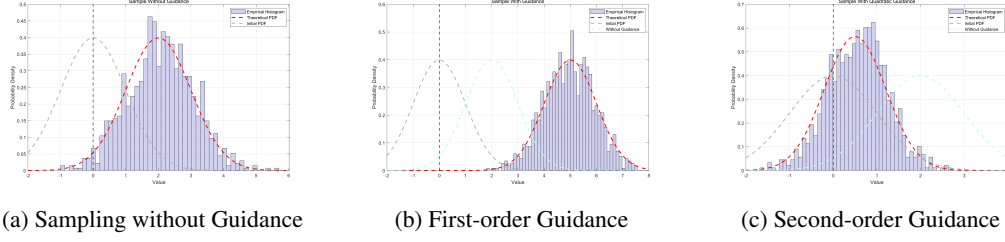


Figure 3: The sampling process of GGDOpt

contrast, with second-order gradient guidance and $\beta = 1$, the distribution is modified to $\mathcal{N}(1/2, 1/2)$, reflecting changes in both mean and variance. These results confirm that GGDOpt effectively directs the sampling process to the desired end distribution. Furthermore, setting $T = 1000$ is sufficient to eliminate the limited time length error.

C.3 Additional experimental results

C.3.1 Linear chance constrained problem

Consider the following linear chance constrained problem

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & \frac{1}{2} \mathbf{x}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{x} \\ \text{s.t.} \quad & \text{Prob}_{\mathbf{c} \sim p_{\mathbf{c}}} \{\mathbf{c}^\top \mathbf{x} + d \geq 0\} \geq 1 - \rho, \end{aligned} \quad (5)$$

where the uncertain parameter follows a Gaussian distribution $p_{\mathbf{c}} = \mathcal{N}(\mathbf{c}; \bar{\mathbf{c}}, \mathbf{I})$ and the hyperparameters $(\mathbf{b}, \bar{\mathbf{c}}, d, \rho)$ are selected from a predefined test set.

For any $\rho < 0.5$, the linear chance constraint can be expressed as

$$-\Phi^{-1}(\rho) \|\mathbf{x}\|_2 - (\bar{\mathbf{c}}^\top \mathbf{x} + d) \leq 0, \quad (6)$$

where Φ denotes the standard Gaussian cumulative distribution function. Then the linear chance constrained problem (5) can be reformulated as the following second-order cone program:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2} \mathbf{x}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{x} \\ \text{s.t.} \quad & -\Phi^{-1}(\rho) \|\mathbf{x}\|_2 - (\bar{\mathbf{c}}^\top \mathbf{x} + d) \leq 0, \end{aligned} \quad (7)$$

which is solved using CVX (Grant et al. [2008]). In practice, we assume the distribution $p_{\mathbf{c}}$ is unknown and only 100 samples are available. To generate training data, we solve the restricted version of the problem for $N = 1000$ values of z linearly spaced in the interval $[0, 0.5]$.

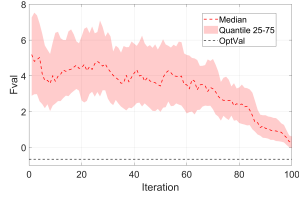
We evaluate the performance of the proposed GGDOpt framework by comparing it with several SAA approaches, using the CVX-based solutions as performance benchmarks. Each algorithm (excluding CVX) is run 100 times and objective values are reported after projecting the solutions onto the feasible set. Experimental results for the case with $n = 8$, $\mathbf{b} = \bar{\mathbf{c}} = (1, 1, \dots, 1)$, $d = 1$, $\rho = 0.1$ are summarized in Table 1, and the sampling process characterized by median and quantiles are provided in Figure 4 to show the stability and fast convergence of GGDOpt.

Based on the results presented in Table 1, we observe that SOC_CVX is capable of exactly identifying the global minimizer of the convexified problem, given full knowledge of the underlying probability distribution. In contrast, SAA-based methods rely solely on sampled realizations and thus yield approximate solutions. Among them, SAA_MIP requires solving a large-scale mixed-integer optimization problem, which is computationally expensive. While SAA_SNSCO demonstrates rapid convergence to optimal solutions in most cases, its performance degrades under worst realizations of \mathbf{h} , occasionally converging to sub-optimal solutions. This leads to strong median performance but instability in statistical results.

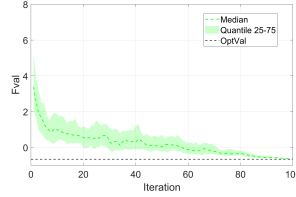
Compared with the SAA methods, our proposed GGDOpt demonstrates superior stability and yields higher-quality solutions, while also significantly reducing computational overhead.

Table 1: Comparison results on the linear chance constrained problem (5)

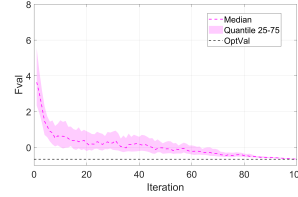
Method	FvalMean	FvalStd	FvalMedian	FvalQuan25	FvalQuan75	Runtime
SOC_CVX (Grant et al. [2008])	-0.6586	0	-0.6586	-0.6586	-0.6586	0.3214
SAA_MIP (Pagnoncelli et al. [2009])	-0.6281	0.0157	-0.6318	-0.6396	-0.6184	15.4502
SAA_CVaR (Nemirovski and Shapiro [2007])	-0.5893	0.0248	-0.5869	-0.6021	-0.5702	0.3063
SAA_SNSCO (Zhou et al. [2024])	0.8051	3.4014	-0.6371	-0.6469	-0.6019	0.2793
SAA_PDCA (Wang et al. [2023])	-0.6389	0.0314	-0.6408	-0.6566	-0.6185	0.6276
GGDOpt (Without Guidance)	0.3481	0.5486	0.2798	-0.0181	0.6142	0.0465
GGDOpt (First-order)	-0.6483	0.0051	-0.6488	-0.6525	-0.6454	0.0486
GGDOpt (Second-order)	-0.6491	0.0056	-0.6503	-0.6531	-0.6474	0.0507



(a) Without Guidance



(b) First-order Guidance



(c) Second-order Guidance

Figure 4: Sampling process visualization of GGDOpt with median and quantiles

165 To provide an intuitive understanding of the sampling behavior in GGDOpt, we illustrate a represen-
 166 tative sampling trajectory of different methods in Figure 5. The results show that, without constraint,
 167 the sampling process will concentrate on the global minimizer of objective function. Under the
 168 influence of constraint, the samples will fall into the feasible set and gradient guidance will lead the
 169 sampling path toward the direction with lower function value. The corresponding iterations of the
 170 objective values for first-order gradient guidance and second-order gradient guidance are shown in
 171 Figure 6.

172 Furthermore, we demonstrate that GGDOpt is capable of producing high-quality solutions across a
 173 range of values for the risk parameter ρ . Specifically, we vary ρ from 0.05 to 0.30 while keeping all
 174 other experimental settings fixed. The corresponding results are reported in Table 2.

175 To further illustrate the efficiency and robustness of GGDOpt, we evaluate its performance under
 176 varying problem dimensions. In particular, we vary the number of decision variables n from 2 to
 177 1024, using the corresponding CVX solutions as performance benchmarks (normalized to 100%).
 178 The comparative performance of GGDOpt under first-order and second-order gradient guidance is
 179 summarized in Table 3.

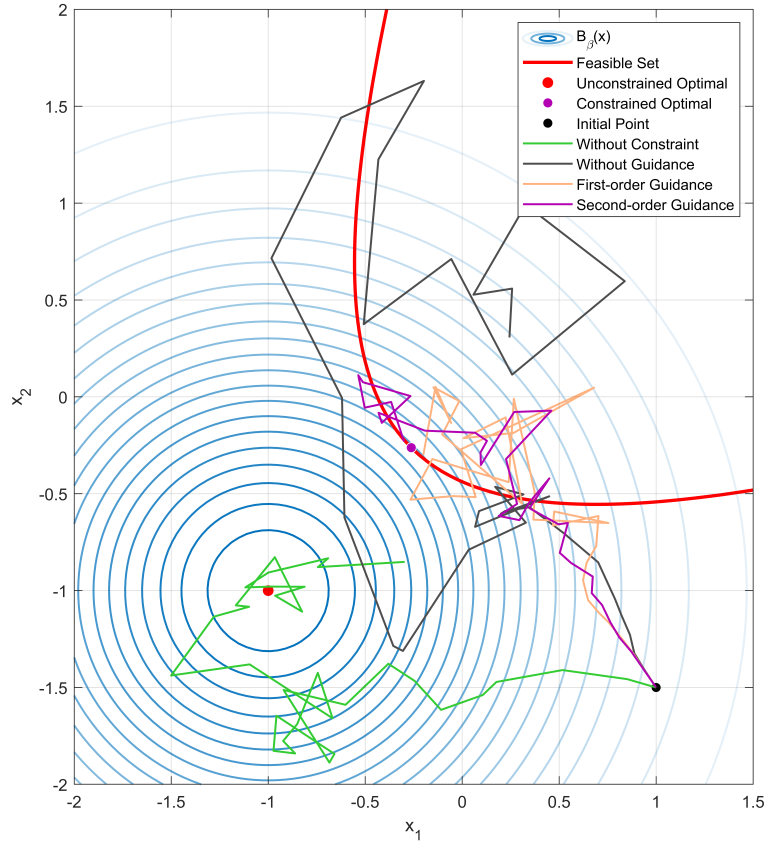


Figure 5: Sampling path of various methods

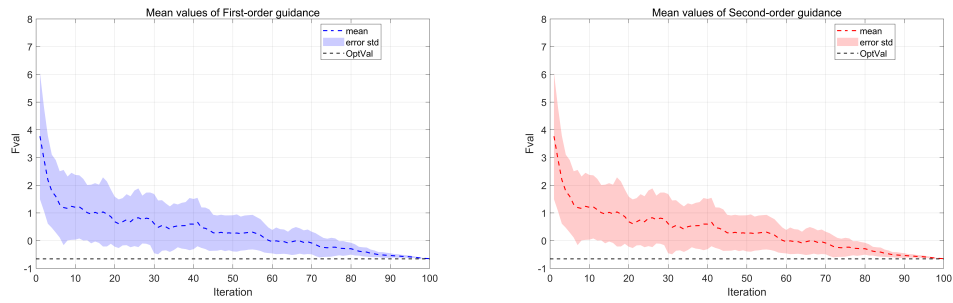


Figure 6: Convergence of the objective values for first- and second-order gradient guidance

Table 2: Comparison results on the linear chance constrained problem (5) with different ρ

Method	$\rho = 0.05$	$\rho = 0.10$	$\rho = 0.15$	$\rho = 0.20$	$\rho = 0.25$	$\rho = 0.30$
SOC_CVX	-0.6073	-0.6585	-0.6983	-0.7335	-0.7667	-0.7991
SAA_MIP	-0.5729	-0.6229	-0.6312	-0.6883	-0.7053	-0.7351
SAA_CVaR	-0.5787	-0.5893	-0.6378	-0.6583	-0.6769	-0.6892
SAA_SNSCO	1.0632	0.8051	-0.2408	-0.4760	-0.5635	-0.6617
SAA_PDCA	-0.5730	-0.6283	-0.6665	-0.6957	-0.7279	-0.7639
GGDOpt_First-order	-0.5955	-0.6483	-0.6828	-0.7032	-0.7284	-0.7603
GGDOpt_Second-order	-0.6040	-0.6491	-0.6944	-0.7130	-0.7498	-0.7817

Table 3: Comparison results on the linear chance constrained problem (5) with different n

Method		SOC_CVX	SAA_PDCA	GGDOpt	GGDOpt
$\rho = 0.1$				(First-order)	(Second-order)
$n = 2$	fval	100.00%	99.39%	99.87%	100.00%
	time	100.00%	111.22%	7.26%	7.31%
$n = 4$	fval	100.00%	96.86%	99.73%	99.89%
	time	100.00%	118.26%	9.86%	10.02%
$n = 8$	fval	100.00%	95.67%	98.44%	98.56%
	time	100.00%	142.25%	15.12%	15.77%
$n = 16$	fval	100.00%	90.42%	98.47%	99.70%
	time	100.00%	160.56%	15.10%	16.06%
$n = 128$	fval	100.00%	95.88%	98.72%	99.89%
	time	100.00%	198.24%	23.63%	25.93%
$n = 1024$	fval	100.00%	93.19%	97.91%	99.34%
	time	100.00%	516.16%	34.11%	37.64%

The results above demonstrate that GGDOpt effectively solves the linear chance constrained problem across varying parameter settings. Moreover, it exhibits significantly higher computational efficiency compared to alternative approaches.

C.3.2 Computational cost

Regarding the computational cost and evaluation, we test the linear chance constrained problem with $n = 8$ and repeat 100 times to calculate the empirical mean of the objective value (fmean), the empirical standard deviation (fstd), and the average run time (time). The results are summarized in the following Table 4.

Table 4: Computational cost of the proposed methods.

Method	SOC_CVX	GGDOpt (First-order)	GGDOpt (Second-order)		
			$\beta = 0.1$	$\beta = 1$	$\beta = 10$
fmean	-0.6586	-0.6483	-0.6341	-0.6548	-0.6585
fstd	0	0.0051	5.6726e-3	2.5112e-05	2.2329e-08
time	0.3214	0.0486	0.0569	0.0527	0.0541

As observed in Table 4, the second-order method achieves lower objective values compared to the first-order method and its performance closely matches the optimal solution obtained by SOC_CVX. Moreover, the second-order method leads to significantly lower standard deviations, particularly as β increases.

We also provide the costs of three stages for the linear chance constraint problem. For each n , we generate 1000 data in the training stage. During sampling, we execute 100 times of reverse process to analyze the stability of GGDOpt. The total time costed in hour is shown in Table 5.

Table 5: Computational time of three stages (in hours).

Stages		$n = 8$	$n = 16$	$n = 128$
Data generating time		0.03	0.06	0.11
Training time		0.53	0.96	11.64
Total sampling time	First-order	0.0013	0.0017	0.0057
	Second-order	0.0014	0.0018	0.0063

Furthermore, our experiments indicate that increasing the quantity of training data alone does not guarantee better performance. Instead, high-quality samples closer to the true optimal solutions are the key of effective guided sampling.

C.3.3 Variance schedule

While Tweedie’s formula theoretically provides both the posterior mean and covariance, $\Sigma_{0|t} = (1 - \bar{\alpha}_t)(\mathbf{I} + (1 - \bar{\alpha}_t)\nabla^2 \log p(\mathbf{x}_t))$, computing the covariance requires evaluating the Hessian of $\log p(\mathbf{x})$.

In our framework, the score function \mathbf{s}_θ is parameterized by a neural network, and computing its second derivatives involves backpropagation through the network’s Jacobian, which is computationally expensive, especially in high dimensions.

To strike a balance between performance and efficiency, we choose to treat the covariance as a tunable constant. This introduces an approximation, but as shown in Table 6, this achieves comparable objective values to the fully Tweedie-based method, while reducing runtime by more than an order of magnitude. These results confirm that using a fixed variance can be a practical and robust alternative.

C.3.4 Guidance term

Our experimental results in Table 7 further demonstrate that for the chance constrained programming, the proposed GGDOpt consistently outperforms the Look-Ahead Guidance from Guo et al. [2024] in terms of both objective value (fval) and computational efficiency (sampling time).

Table 6: Experimental results with different variance schedules.

$n = 8, \rho = 0.1$	Tweedie's Σ	GGDOpt (Second-order)				
		$\sigma = 0.01$	$\sigma = 0.02$	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 10$
fval	-0.6571	-0.6471	-0.6457	-0.6545	-0.6320	-0.6049
time	1.0984	0.0491	0.0496	0.0493	0.0492	0.0493

Table 7: Comparison results with Look-Ahead Guidance Guo et al. [2024].

Method ($\rho = 0.1$)	$n = 2$		$n = 4$		$n = 8$		$n = 16$	
	fval	time	fval	time	fval	time	fval	time
SOC_CVX	-0.4558	0.2148	-0.5630	0.2415	-0.6586	0.3214	-0.7394	0.4067
GGDOpt (First-order)	-0.4552	0.0156	-0.5615	0.0238	-0.6483	0.0486	-0.7281	0.0614
GGDOpt (Second-order)	-0.4558	0.0157	-0.5624	0.0242	-0.6491	0.0507	-0.7372	0.0653
LAG Guo et al. [2024]	-0.4460	0.0329	-0.5181	0.0738	-0.5783	0.1127	-0.6584	0.1436

As shown in the table, our proposed GGDOpt consistently achieves lower objective values and the performance gap between GGDOpt and Look-Ahead Guidance increases with the problem dimension n . In terms of computational efficiency, GGDOpt is approximately $2\times$ faster than the Look-Ahead Guidance across all problem sizes. This performance gain stems from the computational overhead of Guo et al. [2024], where computing the guidance term $G_t^{(3)}$ requires backpropagation through the score network to obtain the gradient of the posterior mean $\mathbb{E}[x_0|x_t]$ with respect to x_t . In contrast, our first- and second-order guidance terms are derived analytically and thus do **not require any additional gradient computations through the network**, making our method more efficient and scalable.

C.3.5 VaR-constrained mean–variance portfolio selection problem

Consider a VaR-constrained mean–variance portfolio selection problem, which aims to minimize the risk while pursuing a targeted level of returns with probability at least $1 - \rho$ (Wang et al. [2023]). Let $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ denote the expectation and covariance matrix of the returns of n risky assets, and $\gamma \in \mathbb{R}_+$ denote the risk aversion factor. Let $x \in \mathbb{R}^n$ denote the allocation vector. Then this problem is formulated as follows:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \gamma x^\top \Sigma x - \mu^\top x \\ \text{s.t.} \quad & \text{Prob}_\xi \{\xi^\top x \geq R\} \geq 1 - \rho, \end{aligned} \tag{8}$$

where $R \in \mathbb{R}_+$ is a prespecified level on the return. We use 2523 daily return data of 435 stocks included in Standard & Poor's 500 Index between March 2006 and March 2016 and set $R = 0.02\%$ and $\gamma = 2$. Some results are shown in Table 8:

In the above experiments, we compare our algorithm with several classical methods, including the mixed-integer program (MIP, Pagnoncelli et al. [2009]), the augmented Lagrangian decomposition method (ALDM, Bai et al. [2021]), the proximal difference-of-convex algorithm (PDCA, Wang et al. [2023]), and the diffusion-based Look-Ahead Guidance (LAG, Guo et al. [2024]) method. We set $\rho = 0.05, 0.1$ and $n = 100, 400$, reporting the final-iteration objective function value (fval), total runtime (time), and the empirical probability of the chance constraint computed over randomly sampled daily returns (prob).

The results show that MIP achieves the lowest objective values but incurs the highest computational cost, as it fully exploits the data by formulating CCP as mixed integer program. LAG attains competitive objectives but requires additional back-propagation steps for guidance. In contrast, GGDOpt well balances solution quality and efficiency, significantly reducing runtime while maintaining comparable objective values and constraint satisfaction.

Table 8: Comparison results of the VaR-constrained mean-variance portfolio selection problem.

(ρ, n)	Metric	MIP	ALDM	PDCA	LAG	GGDOpt (First)	GGDOpt (Second)
(0.05, 100)	fval	-0.0951	-0.0723	-0.0917	-0.0936	-0.0904	-0.0946
	time	15.58	2.418	4.602	0.9433	0.3768	0.4071
	prob	0.8600	0.8666	0.9700	0.8467	0.9200	0.8933
(0.05, 400)	fval	-0.0874	-0.0750	-0.0814	-0.0859	-0.0827	-0.0867
	time	204.2	66.68	93.42	2.7570	1.2732	1.3559
	prob	0.9066	0.8308	0.9891	0.8933	0.9533	0.9267
(0.1, 100)	fval	-0.0951	-0.0721	-0.0856	-0.0927	-0.0915	-0.0936
	time	13.31	2.388	6.258	0.9365	0.3420	0.4218
	prob	0.8600	0.7633	0.9233	0.8533	0.9067	0.8667
(0.1, 400)	fval	-0.0874	-0.0713	-0.0826	-0.0864	-0.0829	-0.0870
	time	148.6	67.95	81.95	2.7323	1.2546	1.2818
	prob	0.9058	0.8158	0.9266	0.8800	0.9267	0.9133

243 C.3.6 Robust waveform design

244 Consider a multiuser multiple-input single-output (MISO) downlink scenario, where a multi-antenna
245 base station transmits independent messages to K single-antenna users over a quasi-static channel.
246 The system model adopted is standard and is briefly described as follows.

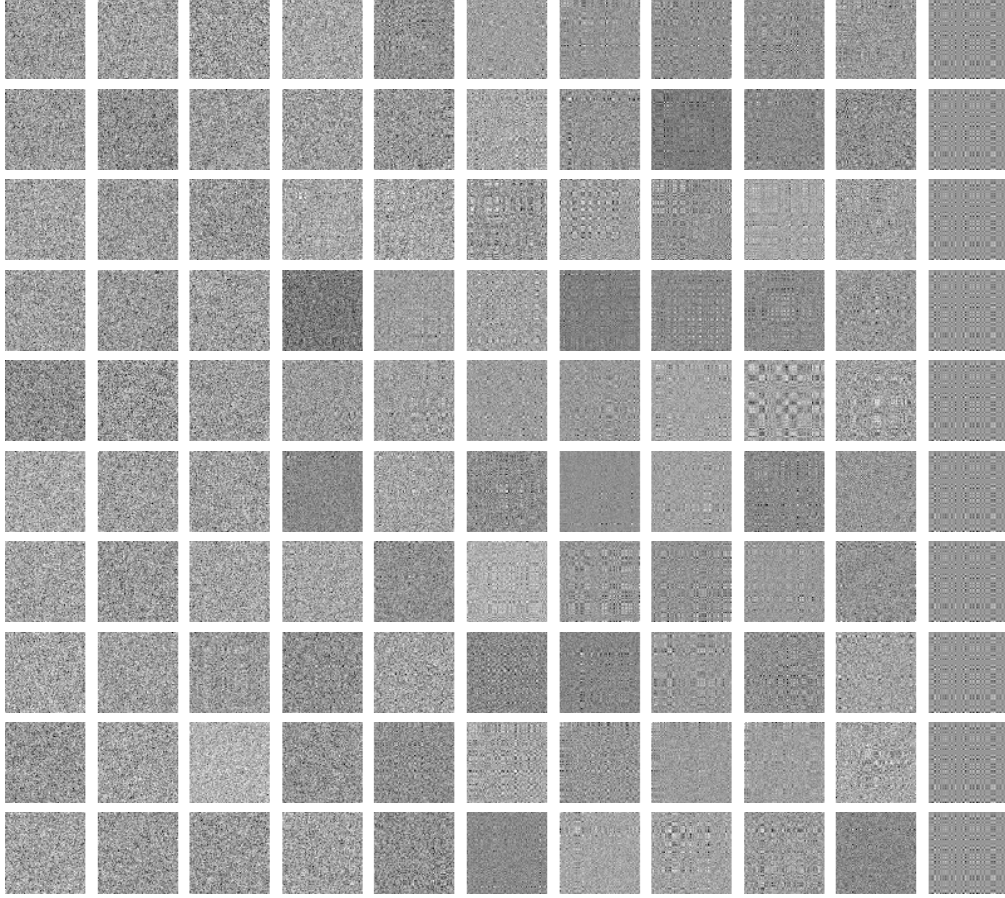


Figure 7: Generated 10 sampling process of GGDOpt with U-Net-2D (from left to right: $t = 100, 90, 80, 70, 60, 50, 40, 30, 20, 10, 0$).

Table 9: Optimization Methods Comparison

$N_t = 64, K = 8$	Metric	$\rho = 0.05$	$\rho = 0.10$	$\rho = 0.15$	$\rho = 0.20$
Empirical Mean	WorstProb	0.4865	0.4865	0.4865	0.4865
	FuncValue	0.0675	0.0675	0.0675	0.0675
	Runtime	4.1406	4.1406	4.1406	4.1406
Sphere Bounding Ben-Tal and Nemirovski [2000]	WorstProb	0.9999	0.9999	0.9999	0.9999
	FuncValue	0.0752	0.0750	0.0749	0.0748
	Runtime	1278	1292	1167	1131
Bernstein-type Inequality Wang et al. [2014]	WorstProb	0.9582	0.9335	0.9122	0.8974
	FuncValue	0.0689	0.0687	0.0686	0.0685
	Runtime	737	703	762	688
GGDOpt (First-order)	WorstProb	0.9521	0.9097	0.8685	0.8107
	FuncValue	0.0692	0.0690	0.0686	0.0685
	Runtime	0.6071	0.5894	0.5941	0.5374
GGDOpt (Second-order)	WorstProb	0.9515	0.9007	0.8573	0.8111
	FuncValue	0.0688	0.0685	0.0684	0.0684
	Runtime	0.6273	0.6152	0.6730	0.5901

Let N_t denote the number of antennae at the base station and K the number of users. The received signal of user $i, i = 1, \dots, K$, is modeled as

$$y_i(t) = \mathbf{h}_i^H \mathbf{x}(t) + \nu_i(t), \quad (9)$$

where $\mathbf{h}_i \in \mathbb{R}^{N_t}$ is the channel of user i ; $\mathbf{x}(t) \in \mathbb{R}^{N_t}$ is the transmit signal from the base station; $\nu_i(t)$ is noise with distribution $\mathcal{N}(0, \sigma_i^2)$.

We assume a general vector-Gaussian linear precoding strategy, where the transmit signal is expressed as

$$\mathbf{x}(t) = \sum_{i=1}^K \mathbf{x}_i(t), \quad (10)$$

with $\mathbf{x}_i(t) \in \mathbb{R}^{N_t}$ representing the information-bearing signal intended for user i . Each $\mathbf{x}_i(t)$ is independently Gaussian encoded with covariance matrix $\mathbf{S}_i \succeq \mathbf{0}$, i.e., $\mathbf{x}_i(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{S}_i)$. At the receiver side, each user decodes only its own intended signal while treating the signals of other users as interference.

Under this system model, the achievable rate for user i can be formulated as

$$R_i = \log_2 \left(1 + \frac{\mathbf{h}_i^H \mathbf{S}_i \mathbf{h}_i}{\sum_{k \neq i} \mathbf{h}_i^H \mathbf{S}_k \mathbf{h}_i + \sigma_i^2} \right), i = 1, \dots, K. \quad (11)$$

To formulate the rate-constrained optimization problem under imperfect channel state information (CSI), it is essential to first characterize the CSI error model. In the presence of imperfect CSI, the actual channel vector of each user can be represented as

$$\mathbf{h}_i = \bar{\mathbf{h}}_i + \mathbf{e}_i, i = 1, \dots, K, \quad (12)$$

where $\bar{\mathbf{h}}_i \in \mathbb{R}^{N_t}$ is the presumed channel at the base station and $\mathbf{e}_i \in \mathbb{R}^{N_t}$ is the channel error vector. We adopt the commonly used Gaussian channel error model. Specifically, each channel error vector is assumed to have a Gaussian distribution, i.e.,

$$\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_i), \quad (13)$$

for some known error covariance matrix C_i . Now, consider the following probabilistically robust design formulation(Wang et al. [2014]):

$$\begin{aligned} \min_{\mathbf{S}_1, \dots, \mathbf{S}_K \in \mathbb{R}^{N_t \times N_t}} \quad & \sum_{i=1}^K \text{Tr}(\mathbf{S}_i) \\ \text{s.t.} \quad & \text{Prob}_{\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)} \{\mathbf{R}_i \geq r_i\} \geq 1 - \rho_i, i = 1, 2, \dots, K, \\ & \mathbf{S}_1, \dots, \mathbf{S}_K \succeq \mathbf{0}, i = 1, 2, \dots, K. \end{aligned} \quad (14)$$

To solve the aforementioned problem using GGDOpt, a naive approach is to treat each covariance matrix as a two-dimensional array and employ a 2D U-Net architecture directly. However, this approach is computationally inefficient, as it requires learning $N_t \times N_t \times K$ variables. To reduce the dimensionality of the optimization variables, we apply Cholesky factorization by expressing each covariance matrix as

$$\mathbf{S}_i = \mathbf{L}_i \mathbf{L}_i^T. \quad (15)$$

This transformation reduces the number of variables per matrix from N_t^2 to $N_t(N_t + 1)/2$, while also ensuring that \mathbf{S}_i remains symmetric and positive semidefinite.

Subsequently, we illustrate representative sampling trajectories of GGDOpt after training (see Figure 7) and observe that the generated solutions consistently approximate rank-one matrices.

Remarkably, the generated samples consistently preserve the rank-one property, with the dominant eigenvalue accounting for over 99% of the total eigenvalue. This observation suggests that solutions to the robust waveform design problem (14) inherently lie on a rank-one manifold with very high probability (Wang et al. [2014]), a structure that GGDOpt can effectively captures. Consequently, rank-one decomposition can be reliably applied after generation, allowing the use of U-Net-1D as a score estimator, which substantially reduces computational costs during both training and sampling process.

Next, we present comparative results for the case $N_t = 64, K = 8$ in Table 9. We compare three approximation methods with our proposed GGDOpt. The Empirical Mean approach directly utilizes the sample mean of the channel realizations $\mathbf{h}_i^{(\ell)}$ and solves the resulting deterministic problem. The Sphere Bounding method (Ben-Tal and Nemirovski [2000]) and the Bernstein-type Inequality approach (Wang et al. [2014]) construct inner convex approximations of the original nonconvex feasible region. For all users, we set $\rho_i = \rho$ for $i = 1, \dots, K$, and evaluate the worst-case outage probability using the true underlying distribution. A solution is deemed feasible if the worst-case probability exceeds $1 - \rho$.

The results demonstrate that across different values of ρ , GGDOpt consistently finds feasible solutions with lower objective values than existing convex restriction methods. Moreover, GGDOpt achieves significantly higher computational efficiency.

By employing U-Net-1D, the sampling process is constrained to produce rank-one solutions. Representative sampling trajectories are illustrated in Figure 8.

D Restricted problem

D.1 Connection with CCP

In this subsection, we establish the connection between the solution of the restricted problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{g}(\mathbf{x}, \bar{\mathbf{h}}) \geq \mathbf{z}, \end{aligned} \quad (16)$$

and that of the CCP

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X}_\rho. \end{aligned} \quad (17)$$

The rationale behind using the restricted problem (RP) to generate high-quality solutions is straightforward. First, solving the restricted problem (16) is computationally more tractable than directly

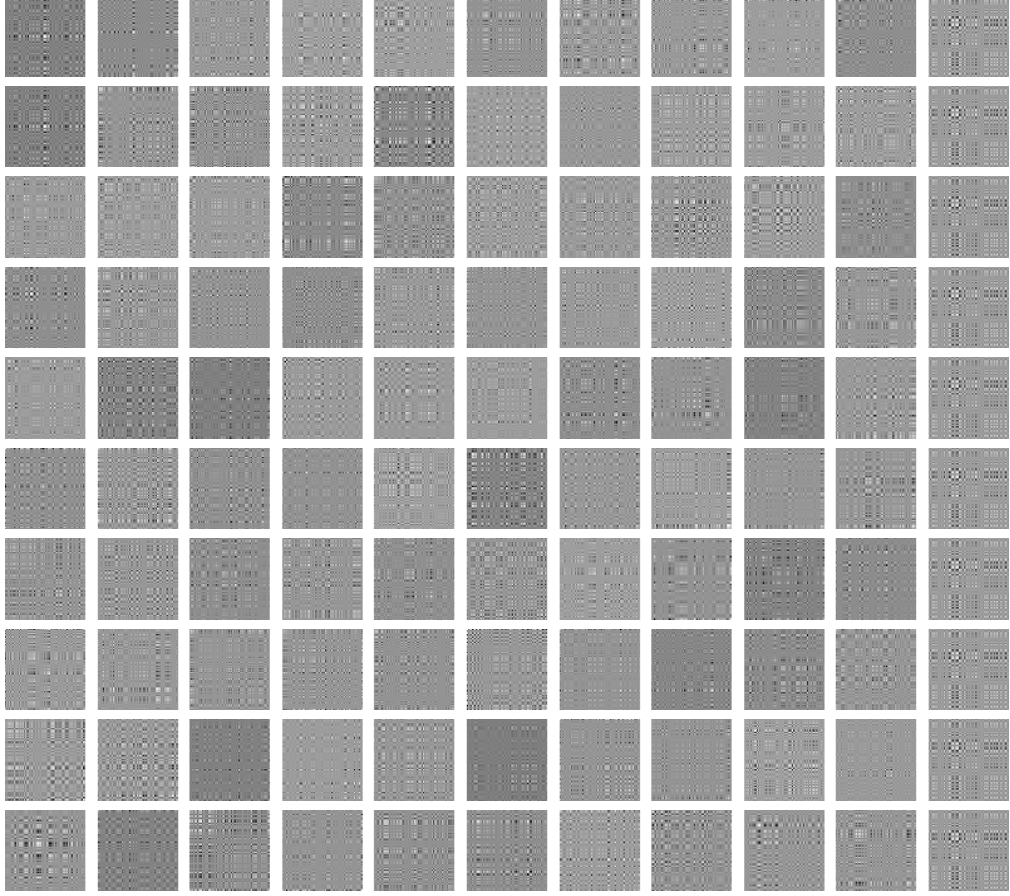


Figure 8: Generated 10 sampling process of GGDOpt with U-Net-1D (from left to right: $t = 100, 90, 80, 70, 60, 50, 40, 30, 20, 10, 0$).

301 tackling the original CCP (17). Second, the distribution P of the random variable \mathbf{h} tends to con-
 302 centrate around its mean $\boldsymbol{\mu}_P$. Consequently, improving the value of $\mathbf{g}(\mathbf{x}, \boldsymbol{\mu}_P)$ generally leads to an
 303 increase in the probability $\text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}, \mathbf{h}) \geq \mathbf{0}\}$. Third, the feasible region of the RP can be viewed
 304 as an approximation of the feasible set \mathcal{X}_ρ associated with the CCP. For a given risk level ρ , solving
 305 the RP yields an approximate local optimum of CCP (17). Moreover, if the global solution to (17)
 306 satisfies certain regularity conditions, this approximate local minimizer coincides with the global
 307 minimizer.

308 In general, the quantity $\text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}(z_i), \mathbf{h}) \geq \mathbf{0}\}$ is hard to compute since it requires a multidimen-
 309 sional integration over the distribution of \mathbf{h} . Inspired by the sample average approximation (SAA),
 310 we estimate this by an empirical average based on L i.i.d. realizations of \mathbf{h} :

$$\text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}(z_i), \mathbf{h}) \geq \mathbf{0}\} \approx \frac{1}{L} \sum_{l=1}^L \ell^{0/1}(\mathbf{g}(\mathbf{x}(z_i), \mathbf{h}^{(l)})) := 1 - \rho^{(i)}, \quad (18)$$

311 where $\ell^{0/1}$ is the element-wise indicator function that returns 1 if all components of the argument
 312 vector are positive, and 0 otherwise. The reason why we choose this to approximate $\rho^{(i)}$ can be
 313 analyzed from the following two situations:

314 On the one hand, if the sample size L is large enough, then the empirical distribution can be regarded
 315 as a good approximation of the underlying distribution, i.e., $p(\mathbf{h}) \approx \frac{1}{L} \sum_{l=1}^L \delta(\mathbf{h} - \mathbf{h}^{(l)})$. In this
 316 case, it is natural to replace the real value that computationally intractable with the empirical value
 317 $\rho^{(i)} = 1 - \frac{1}{L} \sum_{l=1}^L \ell^{0/1}(\mathbf{g}(\mathbf{x}(z_i), \mathbf{h}^{(l)}))$.

On the other hand, if the sample size L is small, using the empirical value to estimate the real ρ will cause serious distortion. In this case, a larger restriction \mathbf{z}_i is preferred, as it will lead to $\mathbf{x}(\mathbf{z}_i)$ with greater probability of satisfying the chance constraint and better robustness to the distribution uncertainty. At this time, the empirical value $\rho^{(i)}$ is not used to approximate the real confidence, but to characterize the properties of "good" $\mathbf{x}(\mathbf{z}_i)$.

To compute $\rho^{(i)}$, we proceed as follows:

- For each sampled restriction vector $\mathbf{z}_i \geq \mathbf{0}$, we solve the corresponding restricted problem, which yields a candidate solution $\mathbf{x}(\mathbf{z}_i)$.
- We then draw L independent realizations $\mathbf{h}^{(\ell)}$ from the underlying distribution and evaluate the fraction of those samples for which $\mathbf{g}(\mathbf{x}(\mathbf{z}_i), \mathbf{h}^{(\ell)}) \geq \mathbf{0}$ holds.

This empirical feasible set constructed in this way provides a conservative inner approximation of the true feasible region, ensuring that the solutions obtained from the restricted problem satisfy the original chance constraint with high confidence.

Next, we provide a detailed characterization of the probability $\mathbf{g}(\mathbf{x}, \mathbf{h}) \geq \mathbf{0}$ evaluated at the solution $\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z})$ to the restricted problem. For brevity, we denote the norm $\|\cdot\| = \|\cdot\|_\infty$ throughout the subsequent analysis.

Assumption 2. Assume that

- (Lipschitz continuity) $\mathbf{g}(\mathbf{x}, \cdot)$ is Lipschitz for a given \mathbf{x} , i.e.,

$$\|\mathbf{g}(\mathbf{x}, \mathbf{h}) - \mathbf{g}(\mathbf{x}, \mathbf{h}')\| \leq L_{\mathbf{x}} \|\mathbf{h} - \mathbf{h}'\|, \quad \forall \mathbf{h}, \mathbf{h}', \quad (19)$$

where $L_{\mathbf{x}}$ is the Lipschitz constant depending on \mathbf{x} .

- (Finite variance) The variance of the random vector \mathbf{h} with probability P is finite, i.e.,

$$\text{Var}_P(\mathbf{h}) < \infty. \quad (20)$$

Theorem 4. Under Assumption 2, suppose that $\{\mathbf{h}^{(\ell)}\}_{\ell=1}^L$ are samples drawn from the distribution P of random vector \mathbf{h} . Let $\bar{\mathbf{h}} = \frac{1}{L} \sum_{\ell=1}^L \mathbf{h}^{(\ell)}$ and let z_{\min} be the smallest element of \mathbf{z} . Suppose that $\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z})$ is the solution to the problem (16), then we have

$$\text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbf{h}) \geq \mathbf{0}\} \geq 1 - \underbrace{\frac{\text{Var}_P(\mathbf{h})}{(z_{\min}/L_{\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z})} - \|\bar{\mathbf{h}} - \mathbb{E}_P[\mathbf{h}]\|)^2}}_{1-\rho}. \quad (21)$$

Proof.

To characterize $\text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbf{h}) \geq \mathbf{0}\}$, we need to consider two sources of error. The first arises from the large variance of the distribution P , while the second stems from the approximation of the mean of P using a finite number of realizations, i.e.,

$$\begin{aligned} & \text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbf{h}) \geq \mathbf{0}\} \\ &= \text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbf{h}) - \mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbb{E}_P[\mathbf{h}]) + \mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbb{E}_P[\mathbf{h}]) - \mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \bar{\mathbf{h}}) \\ & \quad + \mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \bar{\mathbf{h}}) \geq \mathbf{0}\}. \end{aligned} \quad (22)$$

Since $\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \bar{\mathbf{h}})$ is the solution to the restricted problem (16), we have $\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \bar{\mathbf{h}}) \geq z_{\min} \mathbf{1}$. Therefore, we have

$$\begin{aligned} & \text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbf{h}) \geq \mathbf{0}\} \\ & \geq \text{Prob}_{\mathbf{h}}\{\|\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbf{h}) - \mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbb{E}_P[\mathbf{h}])\| + \|\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbb{E}_P[\mathbf{h}]) - \mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \bar{\mathbf{h}})\| \\ & \quad - z_{\min} \leq 0\}. \end{aligned} \quad (23)$$

According to Assumption 2, we have that

$$\|\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbf{h}) - \mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbb{E}_P[\mathbf{h}])\| \leq L_{\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z})} \|\mathbf{h} - \mathbb{E}_P[\mathbf{h}]\|, \quad (24)$$

348 and

$$\|g(x(\bar{h}, z), \mathbb{E}_P[h]) - g(x(\bar{h}, z), \bar{h})\| \leq L_{x(\bar{h}, z)} \|\bar{h} - \mathbb{E}_P[h]\|. \quad (25)$$

349 Therefore, the probability $\text{Prob}_h\{g(x(\bar{h}, z), h) \geq 0\}$ can be further expressed as

$$\begin{aligned} & \text{Prob}_h\{g(x(\bar{h}, z), h) \geq 0\} \\ & \geq \text{Prob}_h\{\|g(x(\bar{h}, z), h) - g(x(\bar{h}, z), \mathbb{E}_P[h])\| \leq z_{\min} - \|g(x(\bar{h}, z), \mathbb{E}_P[h]) - g(x(\bar{h}, z), \bar{h})\|\} \\ & \geq \text{Prob}_h\{L_{x(\bar{h}, z)} \|h - \mathbb{E}_P[h]\| \leq z_{\min} - L_{x(\bar{h}, z)} \|\bar{h} - \mathbb{E}_P[h]\|\} \\ & = \text{Prob}_h\{\|h - \mathbb{E}_P[h]\| \leq z_{\min}/L_{x(\bar{h}, z)} - \|\bar{h} - \mathbb{E}_P[h]\|\}. \end{aligned} \quad (26)$$

350 By Chebyshev's inequality (Chebyshev [1867]), we obtain that

$$\begin{aligned} & \text{Prob}_h\{\|h - \mathbb{E}_P[h]\| \leq z_{\min}/L_{x(\bar{h}, z)} - \|\bar{h} - \mathbb{E}_P[h]\|\} \\ & \geq 1 - \frac{\text{Var}_P(h)}{(z_{\min}/L_{x(\bar{h}, z)} - \|\bar{h} - \mathbb{E}_P[h]\|)^2}. \end{aligned} \quad (27)$$

351 Hence, we have

$$\text{Prob}_h\{g(x(\bar{h}, z), h) \geq 0\} \geq 1 - \frac{\text{Var}_P(h)}{(z_{\min}/L_{x(\bar{h}, z)} - \|\bar{h} - \mathbb{E}_P[h]\|)^2}. \quad (28)$$

352

□

353 Theorem 4 demonstrates that as z_{\min} increases, the lower bound on the probability that the chance
 354 constraint is satisfied at the point $x(\bar{h}, z)$ also increases. This implies that $x(\bar{h}, z)$ is more likely to
 355 be a feasible solution to the CCP (17), while potentially achieving a lower objective value. In the
 356 following theorem, we further establish that, under certain regularity conditions, the global minimizer
 357 of the CCP (17) is contained within the set of solutions to the restricted problem (16).

358 **Assumption 3.** Assume that

359 • (Bounded bias) For any given ρ , denote $x^* = \arg \min_{x \in \mathcal{X}_\rho} f(x)$, then

$$\|\bar{h} - \mathbb{E}_P[h]\| \leq \frac{g(x^*, \bar{h})}{L_{x(\bar{h}, z)}} - \sqrt{\frac{\text{Var}_P(h)}{\rho}}. \quad (29)$$

360 • (Reliable data set) For the generated data set $\mathcal{D} = \{(x^{(i)}, \rho^{(i)})\}_{i=1}^N$, $\rho^{(i)}$ is a lower bound of real
 361 probability $\text{Prob}_h\{g(x^{(i)}, h) \geq 0\}$.

362 Note that Assumption 3 can be satisfied with a sufficiently large number of realizations of h and the
 363 corresponding restriction estimator. For instance, we can choose

$$\rho^{(i)} \leq \frac{\text{Var}_P(h)}{(z_{\min}/L_{x(\bar{h}, z)} - \|\bar{h} - \mathbb{E}_P[h]\|)^2}. \quad (30)$$

364 **Theorem 5.** Under Assumption 2 and Assumption 3, for any given ρ and \bar{h} , suppose that

$$\mathcal{D}_\rho = \{x^{(i)} \mid (x^{(i)}, \rho^{(i)}) \in \mathcal{D}, \rho^{(i)} \leq \rho\}, \quad (31)$$

365 then we have

$$x^* \in \mathcal{D}_\rho \subset \mathcal{X}_\rho. \quad (32)$$

366 *Proof.*

367 We choose z_{\min} as the smallest element of $g(x^*, \bar{h})$, then for any x that satisfies $g(x, \bar{h}) \geq z$, the
 368 following inequality holds:

$$\text{Prob}_h\{g(x, h) \geq 0\} \geq 1 - \frac{\text{Var}_P(h)}{(z_{\min}/L_{x(\bar{h}, z)} - \|\bar{h} - \mathbb{E}_P[h]\|)^2} \geq 1 - \rho. \quad (33)$$

369 This implies that

$$\{\mathbf{x} \mid \mathbf{g}(\mathbf{x}, \bar{\mathbf{h}}) \geq \mathbf{z}\} \subset \mathcal{X}_\rho. \quad (34)$$

370 Recall the definition of $\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z})$, which is the global minimizer of $f(\mathbf{x})$ over the set $\{\mathbf{x} \mid \mathbf{g}(\mathbf{x}, \bar{\mathbf{h}}) \geq \mathbf{z}\}$.
 371 Additionally, it follows naturally that $\mathbf{g}(\mathbf{x}^*, \bar{\mathbf{h}}) \geq \mathbf{z}$, i.e.,

$$\mathbf{x}^* \in \{\mathbf{x} \mid \mathbf{g}(\mathbf{x}, \bar{\mathbf{h}}) \geq \mathbf{z}\}. \quad (35)$$

372 This implies that \mathbf{x}^* is also a global minimizer of $f(\mathbf{x})$ over the set $\{\mathbf{x} \mid \mathbf{g}(\mathbf{x}, \bar{\mathbf{h}}) \geq \mathbf{z}\}$. Therefore,
 373 we have

$$\mathbf{x}^* \in \mathcal{D}_\rho \subset \mathcal{X}_\rho. \quad (36)$$

374

□

375 This result plays a crucial role in the GGDOpt framework, as the sampler is inherently limited to
 376 generating solutions that are no better than the quality of the training data. Theoretical guarantees
 377 established above indicate that the data generated from the restricted problem are sufficiently infor-
 378 mative and may contain the true global minimizer of the CCP (17). This justifies the effectiveness of
 379 using such data to train our GGDOpt.

380 D.2 Special cases

381 The above results provide a lower bound for the probability $\text{Prob}_{\mathbf{h}}\{\mathbf{g}(\mathbf{x}(\bar{\mathbf{h}}, \mathbf{z}), \mathbf{h}) \geq 0\}$. In most
 382 cases, the explicit value of this probability cannot be directly computed. However, in this subsection,
 383 we present a special case corresponding to the robust waveform design problem, where the probability
 384 can be expressed explicitly.

385 **Theorem 6.** Suppose $\mathbf{x}^*(\bar{\mathbf{h}}_i, \mathbf{z})$ is the solution to the following restricted problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}, \bar{\mathbf{h}}_i) = z_i, i = 1, \dots, K, \end{aligned} \quad (37)$$

386 where $g_i(\mathbf{x}, \cdot)$ is a quadratic function of \mathbf{h} with parameters $(\mathbf{A}_i, \mathbf{b}_i, d_i)$ and the parameters $\mathbf{h}_i \sim$
 387 $\mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)$. Denote

$$\begin{aligned} \mathbf{Q}_i &= \mathbf{C}_i^{1/2} \mathbf{A}_i \mathbf{C}_i^{1/2} \stackrel{\text{svd}}{=} \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T, \\ \mathbf{r}_i &= \mathbf{C}_i^{1/2} (\mathbf{A}_i \bar{\mathbf{h}}_i + \mathbf{b}_i), \\ s_i &= \frac{1}{2} \bar{\mathbf{h}}_i^\top \mathbf{A}_i \bar{\mathbf{h}}_i + \mathbf{b}_i^\top \bar{\mathbf{h}}_i + d_i, \\ \mathbf{c}_i &= \mathbf{U}_i^T \mathbf{r}_i, \end{aligned} \quad (38)$$

388 and let

$$\begin{aligned} \mathbf{u}_i &= \mathbf{U}_i^T \mathbf{e}_i, \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ Y_i &= \frac{1}{2} \mathbf{u}_i^\top \mathbf{\Lambda}_i \mathbf{u}_i + \mathbf{c}_i^\top \mathbf{u}_i + s_i. \end{aligned} \quad (39)$$

389 Then for $\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)$, we have

$$\text{Prob}_{\mathbf{h}_i}\{g_i(\mathbf{x}^*, \mathbf{h}_i) \geq 0\} = 1 - F_{Y_i}(0), \quad (40)$$

390 where F_{Y_i} is the cumulative distribution function of Y_i .

391 *Proof.*

392 For quadratic $g_i(\mathbf{x}, \cdot)$ of \mathbf{h} with parameters $(\mathbf{A}_i, \mathbf{b}_i, d_i)$ and given that $\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)$, the proba-
 393 bility $\text{Prob}_{\mathbf{h}_i}\{g_i(\mathbf{x}, \mathbf{h}_i) \geq 0\}$ can be transformed into the following form:

$$\begin{aligned} & \text{Prob}_{\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)}\{g_i(\mathbf{x}, \mathbf{h}_i) \geq 0\} \\ &= \text{Prob}_{\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)}\left\{\frac{1}{2} \mathbf{h}_i^\top \mathbf{A}_i \mathbf{h}_i + \mathbf{b}_i^\top \mathbf{h}_i + d_i \geq 0\right\} \\ &= \text{Prob}_{\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\left\{\frac{1}{2} (\bar{\mathbf{h}}_i + \mathbf{C}_i^{1/2} \mathbf{e}_i)^\top \mathbf{A}_i (\bar{\mathbf{h}}_i + \mathbf{C}_i^{1/2} \mathbf{e}_i) + \mathbf{b}_i^\top (\bar{\mathbf{h}}_i + \mathbf{C}_i^{1/2} \mathbf{e}_i) + d_i \geq 0\right\}. \end{aligned} \quad (41)$$

394 Denote

$$\begin{aligned}\mathbf{Q}_i &= \mathbf{C}_i^{1/2} \mathbf{A}_i \mathbf{C}_i^{1/2} \stackrel{\text{svd}}{=} \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T, \\ \mathbf{r}_i &= \mathbf{C}_i^{1/2} (\mathbf{A}_i \bar{\mathbf{h}}_i + \mathbf{b}_i), \\ s_i &= \frac{1}{2} \bar{\mathbf{h}}_i^\top \mathbf{A}_i \bar{\mathbf{h}}_i + \mathbf{b}_i^\top \bar{\mathbf{h}}_i + d_i,\end{aligned}\tag{42}$$

395 then we have

$$\text{Prob}_{\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)} \{g_i(\mathbf{x}, \mathbf{h}_i) \geq 0\} = \text{Prob}_{\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\{ \frac{1}{2} \mathbf{e}_i^\top \mathbf{Q}_i \mathbf{e}_i + \mathbf{r}_i^\top \mathbf{e}_i + s_i \geq 0 \right\}.\tag{43}$$

396 Denote $\mathbf{Q}_i \stackrel{\text{svd}}{=} \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T$ and let

$$\begin{aligned}\mathbf{c}_i &= \mathbf{U}_i^T \mathbf{r}_i, \\ \mathbf{u}_i &= \mathbf{U}_i^T \mathbf{e}_i, \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ Y_i &= \frac{1}{2} \mathbf{u}_i^\top \mathbf{\Lambda}_i \mathbf{u}_i + \mathbf{c}_i^\top \mathbf{u}_i + s_i.\end{aligned}\tag{44}$$

397 Substituting these expressions into the above probability, we obtain that

$$\begin{aligned}& \text{Prob}_{\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)} \{g_i(\mathbf{x}, \mathbf{h}_i) \geq 0\} \\ &= \text{Prob}_{\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\{ \frac{1}{2} \mathbf{u}_i^\top \mathbf{\Lambda}_i \mathbf{u}_i + \mathbf{c}_i^\top \mathbf{u}_i + s_i \geq 0 \right\} \\ &= \text{Prob}_{\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \{Y_i \geq 0\}.\end{aligned}\tag{45}$$

398 Denote $\lambda_i^{(k)}$, $u_i^{(k)}$, and $c_i^{(k)}$ as the k -th element of $\mathbf{\Lambda}_i$, \mathbf{u}_i , and \mathbf{c}_i , where $k = 1, \dots, n$. Note that Y_i
399 has a quadratic form of standard Gaussian \mathbf{u}_i , which can be reformulated as a standard quadratic
400 form:

$$Y_i = \sum_{\lambda_i^{(k)} \neq 0} \frac{\lambda_i^{(k)}}{2} \left(u_i^{(k)} + \frac{c_i^{(k)}}{\lambda_i^{(k)}} \right)^2 + \sum_{\lambda_i^{(k)} = 0} c_i^{(k)} u_i^{(k)} + \left(s_i - \sum_{\lambda_i^{(k)} \neq 0} \frac{(c_i^{(k)})^2}{2\lambda_i^{(k)}} \right),\tag{46}$$

401 where $\left(u_i^{(k)} + \frac{c_i^{(k)}}{\lambda_i^{(k)}} \right)^2 \sim \chi_1^2 \left(\left(\frac{c_i^{(k)}}{\lambda_i^{(k)}} \right)^2 \right)$ follows noncentral chi-squared distribution and $c_i^{(k)} u_i^{(k)} \sim$
402 $\mathcal{N}(0, (c_i^{(k)})^2)$ follows Gaussian distribution.

403 Denote F_{Y_i} as the cumulative distribution function of Y_i , then we have

$$\text{Prob}_{\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)} \{g_i(\mathbf{x}, \mathbf{h}_i) \geq 0\} = 1 - F_{Y_i}(0).\tag{47}$$

404 Since $\mathbf{x}^*(\bar{\mathbf{h}}_i, \mathbf{z})$ is the solution to the restricted problem, by substituting $s_i = z_i$, we obtain the result
405 of Theorem 6.

406 □

407 Theorem 6 tells us that the probability $\text{Prob}_{\mathbf{h}_i} \{g_i(\mathbf{x}^*, \mathbf{h}_i) \geq 0\}$ can be expressed in terms of the
408 cumulative distribution function of Y_i . Note that Y_i consists of n independent variables. The following
409 corollary states that, for sufficiently large n , Y_i can be approximated as a Gaussian random variable,
410 and the probability can be computed using the standard Gaussian cumulative distribution function Φ .

411 **Corollary 2.** For sufficiently large n , the probability can be approximated by

$$\text{Prob}_{\mathbf{h}_i} \{g_i(\mathbf{x}, \mathbf{h}_i) \geq 0\} \approx 1 - \Phi \left(\frac{-\mu_{Y_i}}{\sigma_{Y_i}} \right),\tag{48}$$

412 where Φ denotes the cumulative distribution function of the standard Gaussian distribution and

$$\begin{aligned}\mu_{Y_i} &= \frac{1}{2} \text{tr}(\mathbf{Q}_i) + z_i, \\ \sigma_{Y_i}^2 &= \frac{1}{2} \|\mathbf{Q}_i\|_F^2 + \|\mathbf{r}_i\|^2.\end{aligned}\tag{49}$$

413 The approximation error can be bounded by

$$\left| F_{Y_i}(0) - \Phi\left(\frac{-\mu_{Y_i}}{\sigma_{Y_i}}\right) \right| = O(n^{-1/2}). \quad (50)$$

414 *Proof.*

415 For sufficiently large n , the distribution of Y_i can be approximated by Gaussian distribution
416 $\mathcal{N}(\mu_{Y_i}, \sigma_{Y_i}^2)$ with central limit theorem, where

$$\begin{aligned} \mu_{Y_i} &= \frac{1}{2} \text{tr}(\mathbf{Q}_i) + z_i, \\ \sigma_{Y_i}^2 &= \frac{1}{2} \|\mathbf{Q}_i\|_F^2 + \|\mathbf{r}_i\|^2, \end{aligned} \quad (51)$$

417 then the probability can be approximated by

$$\text{Prob}_{\mathbf{h}_i \sim \mathcal{N}(\bar{\mathbf{h}}_i, \mathbf{C}_i)} \{g_i(\mathbf{x}, \mathbf{h}_i) \geq 0\} \approx 1 - \Phi\left(\frac{-\mu_{Y_i}}{\sigma_{Y_i}}\right). \quad (52)$$

418 The approximation error can be bounded by Klartag and Sodin [2012]

$$|F_{Y_i}(0) - \Phi\left(\frac{-\mu_{Y_i}}{\sigma_{Y_i}}\right)| = O(n^{-1/2}). \quad (53)$$

419

□

420 E Technical appendices

421 E.1 Proof of Theorem 1

422 **Theorem 1.** For any given $\beta > 0$, there exists $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ such that the score function of the diffused
423 product distribution can be formulated as

$$\nabla_{\mathbf{x}_t} \log \tilde{p}_t(\mathbf{x}_t|\rho) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\rho) - \underbrace{\beta \nabla_{\mathbf{x}_t} f(\hat{\mathbf{x}}_0(\mathbf{x}_t))}_{\text{gradient guidance } \mathbf{G}_t}, \quad (54)$$

424 where $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\rho)$ is the score function of the diffused data distribution and $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ satisfies

$$f(\hat{\mathbf{x}}_0(\mathbf{x}_t)) = -\frac{1}{\beta} \log \left(\int_{\mathbf{x}_0} p_{t0}(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0 \right). \quad (55)$$

425 *Proof.*

426 Given $\tilde{p}_0(\mathbf{x}_0|\rho)$ and the forward process $d\mathbf{x} = \mathbf{a}(\mathbf{x}, t)dt + b(t)d\mathbf{B}_t$, the diffused conditional distri-
427 bution of unguided distribution $p_0(\mathbf{x}_0|\rho)$ and product distribution $\tilde{p}_0(\mathbf{x}_0|\rho)$ satisfies

$$\begin{aligned} p_t(\mathbf{x}_t|\rho) &= \int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0) p_0(\mathbf{x}_0|\rho) d\mathbf{x}_0, \\ \tilde{p}_t(\mathbf{x}_t|\rho) &= \int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0) \tilde{p}_0(\mathbf{x}_0|\rho) d\mathbf{x}_0 \propto \int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0) p_0(\mathbf{x}_0|\rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0. \end{aligned} \quad (56)$$

428 Consider the difference between the score function of unguided $p_t(\mathbf{x}_t|\rho)$ and guided $\tilde{p}_t(\mathbf{x}_t|\rho)$, we
429 have that

$$\begin{aligned} &\nabla_{\mathbf{x}_t} \log \tilde{p}_t(\mathbf{x}_t|\rho) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\rho) \\ &= \nabla_{\mathbf{x}_t} \log \int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0) p_0(\mathbf{x}_0|\rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0 - \nabla_{\mathbf{x}_t} \log \int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0) p_0(\mathbf{x}_0|\rho) \\ &= \nabla_{\mathbf{x}_t} \log \frac{\int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0) p_0(\mathbf{x}_0|\rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0}{\int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0) p_0(\mathbf{x}_0|\rho)}. \end{aligned} \quad (57)$$

430 Notice that the inner fractional part can be expressed by

$$\frac{p_{0t}(\mathbf{x}_t|\mathbf{x}_0)p_0(\mathbf{x}_0|\rho)}{\int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)p_0(\mathbf{x}_0|\rho)} = p(\mathbf{x}_0|\mathbf{x}_t, \rho), \quad (58)$$

431 then we have

$$\nabla_{\mathbf{x}_t} \log \tilde{p}_t(\mathbf{x}_t|\rho) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\rho) = \nabla_{\mathbf{x}_t} \log \int_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0. \quad (59)$$

432 One way to tackle the log integral is to use the mean value theorem. There exists $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ such that

$$\int_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0 = B_\beta(\hat{\mathbf{x}}_0(\mathbf{x}_t)) \int_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{x}_t, \rho) d\mathbf{x}_0. \quad (60)$$

433 Then we have

$$\nabla_{\mathbf{x}_t} \log \tilde{p}_t(\mathbf{x}_t|\rho) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\rho) = \nabla_{\mathbf{x}_t} \log B_\beta(\hat{\mathbf{x}}_0(\mathbf{x}_t)) = -\beta \nabla_{\mathbf{x}_t} f(\hat{\mathbf{x}}_0(\mathbf{x}_t)), \quad (61)$$

434 and $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ satisfies

$$f(\hat{\mathbf{x}}_0(\mathbf{x}_t)) = -\frac{1}{\beta} \log \left(\frac{\int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)p_0(\mathbf{x}_0|\rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0}{\int_{\mathbf{x}_0} p_{0t}(\mathbf{x}_t|\mathbf{x}_0)p_0(\mathbf{x}_0|\rho) d\mathbf{x}_0} \right). \quad (62)$$

435

□

436 E.2 Proof of Corollary 1

437 **Corollary 1.** Assume that $p_{t0}(\mathbf{x}_0|\mathbf{x}_t, \rho) = \mathcal{N}(\mathbf{x}_0|\boldsymbol{\mu}_{0|t}, \sigma_{0|t}^2 \mathbf{I})$, then we have the following results.

438 • **First-order guidance:** For $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$, we get

$$\mathbf{G}_t = -\beta \nabla_{\mathbf{x}_t} f(\mathbf{x}_t). \quad (63)$$

439 • **Second-order guidance:** For $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$, we get

$$\mathbf{G}_t = -\frac{1}{\sigma_{0|t}^2} \left[\mathbf{H}^{-1} \left((-\nabla_{\mathbf{x}_t}^2 f(\mathbf{x}_t) \mathbf{x}_t + \nabla_{\mathbf{x}_t} f(\mathbf{x}_t)) - \frac{1}{\beta \sigma_{0|t}^2} \boldsymbol{\mu}_{0|t} \right) + \boldsymbol{\mu}_{0|t} \right], \quad (64)$$

440 where $\mathbf{H} = \nabla_{\mathbf{x}_t}^2 f(\mathbf{x}_t) + \frac{1}{\beta \sigma_{0|t}^2} \mathbf{I}$.

441 *Proof.*

442 Due to the implicit nature of $\hat{\mathbf{x}}_0(\mathbf{x}_t)$, directly computing $\nabla_{\mathbf{x}_t} f(\hat{\mathbf{x}}_0(\mathbf{x}_t))$ is intractable. Therefore,
443 we consider an alternative approach by directly examining $\nabla_{\mathbf{x}_t} f(\hat{\mathbf{x}}_0(\mathbf{x}_t))$. By performing the
444 differentiation $\nabla_{\mathbf{x}_t}$, we obtain

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log \tilde{p}_t(\mathbf{x}_t|\rho) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\rho) &= \nabla_{\mathbf{x}_t} \log \int_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0 \\ &= \frac{\int_{\mathbf{x}_0} \nabla_{\mathbf{x}_t} p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0}{\int_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0}. \end{aligned} \quad (65)$$

445 According to the assumption that $p_{t0}(\mathbf{x}_0|\mathbf{x}_t, \rho) = \mathcal{N}(\mathbf{x}_0|\boldsymbol{\mu}_{0|t}, \sigma_{0|t}^2 \mathbf{I})$, we have

$$\nabla_{\mathbf{x}_t} p(\mathbf{x}_0|\mathbf{x}_t, \rho) = \frac{\mathbf{x}_0 - \boldsymbol{\mu}_{0|t}}{\sigma_{0|t}^2} p(\mathbf{x}_0|\mathbf{x}_t, \rho). \quad (66)$$

446 Substituting into the above result, we have

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log \tilde{p}_t(\mathbf{x}_t|\rho) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\rho) &= \frac{\int_{\mathbf{x}_0} \frac{\mathbf{x}_0 - \boldsymbol{\mu}_{0|t}}{\sigma_{0|t}^2} p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0}{\int_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0} \\ &= \frac{1}{\sigma_{0|t}^2} \left(\frac{\int_{\mathbf{x}_0} \mathbf{x}_0 p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0}{\int_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{x}_t, \rho) B_\beta(\mathbf{x}_0) d\mathbf{x}_0} - \boldsymbol{\mu}_{0|t} \right) \\ &= \frac{1}{\sigma_{0|t}^2} (\mathbb{E}[\tilde{\mathbf{x}}] - \boldsymbol{\mu}_{0|t}), \end{aligned} \quad (67)$$

447 where $\tilde{\mathbf{x}} \sim p(\tilde{\mathbf{x}}) \propto p(\mathbf{x}_0|\mathbf{x}_t, \rho)B_\beta(\mathbf{x}_0)$. Given an objective f with the following quadratic form:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}, \quad (68)$$

448 we have

$$B_\beta(\mathbf{x}_0) \propto e^{-\beta f(\mathbf{x})} = e^{-\beta(\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x})}. \quad (69)$$

449 For $\beta\mathbf{A} + \frac{1}{\sigma_{0|t}^2}\mathbf{I} \succ \beta\mathbf{A} + \frac{1}{\sigma_0^2}\mathbf{I} \succ 0$, we have

$$\mathbb{E}[\tilde{\mathbf{x}}] = -\left(\beta\mathbf{A} + \frac{1}{\sigma_{0|t}^2}\mathbf{I}\right)^{-1}\left(\beta\mathbf{b} - \frac{1}{\sigma_{0|t}^2}\boldsymbol{\mu}_{0|t}\right), \quad (70)$$

450 and then we have gradient guidance

$$\begin{aligned} \mathbf{G}_t &= \nabla_{\mathbf{x}_t} \log \tilde{p}_t(\mathbf{x}_t|\rho) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\rho) \\ &= -\frac{1}{\sigma_{0|t}^2} \left[\left(\beta\mathbf{A} + \frac{1}{\sigma_{0|t}^2}\mathbf{I}\right)^{-1}(\beta\mathbf{b} - \frac{1}{\sigma_{0|t}^2}\boldsymbol{\mu}_{0|t}) + \boldsymbol{\mu}_{0|t} \right]. \end{aligned} \quad (71)$$

451 For a general objective f , if we use the first-order Taylor expansion

$$f(\mathbf{x}) \approx f(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} f(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t), \quad (72)$$

452 then the Gradient Guidance can be formulated as the following form by setting $\mathbf{A} = \mathbf{0}, \mathbf{b} =$
453 $\nabla_{\mathbf{x}_t} f(\mathbf{x}_t)$:

$$\mathbf{G}_t = -\beta \nabla_{\mathbf{x}_t} f(\mathbf{x}_t). \quad (73)$$

454 If we use the second-order Taylor expansion

$$f(\mathbf{x}) \approx f(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} f(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top \nabla_{\mathbf{x}_t}^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t), \quad (74)$$

455 then the Gradient Guidance can be formulated as the following form by setting $\mathbf{A} = \nabla_{\mathbf{x}_t}^2 f(\mathbf{x}_t), \mathbf{b} =$
456 $-\nabla_{\mathbf{x}_t}^2 f(\mathbf{x}_t)\mathbf{x}_t + \nabla_{\mathbf{x}_t} f(\mathbf{x}_t)$:

$$\mathbf{G}_t = -\frac{1}{\sigma_{0|t}^2} \left[\left(\beta \nabla_{\mathbf{x}_t}^2 f(\mathbf{x}_t) + \frac{1}{\sigma_{0|t}^2}\mathbf{I}\right)^{-1} \left[\beta(-\nabla_{\mathbf{x}_t}^2 f(\mathbf{x}_t)\mathbf{x}_t + \nabla_{\mathbf{x}_t} f(\mathbf{x}_t)) - \frac{1}{\sigma_{0|t}^2}\boldsymbol{\mu}_{0|t} \right] + \boldsymbol{\mu}_{0|t} \right]. \quad (75)$$

457

□

458 The posterior assumption in Corollary 1 can be satisfied easily. For example, with $p_0(\mathbf{x}_0|\rho) =$
459 $\mathcal{N}(\mathbf{x}_0|\boldsymbol{\mu}_0, \sigma_0^2\mathbf{I})$ and forward process

$$d\mathbf{x} = -\theta\mathbf{x}dt + \sqrt{2\theta}d\mathbf{B}_t, \quad (76)$$

460 we have

$$p_t(\mathbf{x}_t|\rho) = \mathcal{N}\left(\mathbf{x}_t|\boldsymbol{\mu}_0 e^{-\theta t}, (\sigma_0^2 e^{-2\theta t} + 1 - e^{-2\theta t})\mathbf{I}\right). \quad (77)$$

461 Denote $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 e^{-\theta t}, \sigma_t^2 = \sigma_0^2 e^{-2\theta t} + 1 - e^{-2\theta t}$, we have

$$p(\mathbf{x}_0|\mathbf{x}_t, \rho) = \mathcal{N}(\mathbf{x}_0|\boldsymbol{\mu}_{0|t}, \sigma_{0|t}^2\mathbf{I}), \quad (78)$$

462 where

$$\begin{aligned} \boldsymbol{\mu}_{0|t} &= \boldsymbol{\mu}_0 + \frac{\sigma_0^2}{\sigma_t^2} e^{-\theta t} (\mathbf{x}_t - \boldsymbol{\mu}_t), \\ \sigma_{0|t}^2 &= \sigma_0^2 \left(1 - \frac{\sigma_0^2}{\sigma_t^2} e^{-2\theta t} \right). \end{aligned} \quad (79)$$

463 E.3 Proof of Theorem 2

464 **Assumption 4.** For the forward process

$$d\mathbf{x}_t = \mathbf{a}(\mathbf{x}_t, t)dt + b(t)d\mathbf{B}_t, \quad (80)$$

465 there is a constant C such that

- 466 (i) $\mathbf{a}(\mathbf{x}_t, t)$ is globally Lipschitz for any $t \in [0, T]$, i.e. $\|\mathbf{a}(\mathbf{x}_t, t) - \mathbf{a}(\mathbf{x}'_t, t)\| \leq C\|\mathbf{x} - \mathbf{x}'_t\|$;
- 467 (ii) $\mathbf{a}(\mathbf{x}_t, t)$ grows at most linearly for any $t \in [0, T]$, i.e. $\|\mathbf{a}(\mathbf{x}_t, t)\| \leq C(1 + \|\mathbf{x}_t\|)$;
- 468 (iii) \mathbf{x}_t has a density $p_t \in \mathcal{C}^1$ for every $t > 0$ and

$$\int_{t_0}^1 \int_{\|\mathbf{x}_t\| < R} |p_t(\mathbf{x}_t)|^2 + \|\nabla_{\mathbf{x}_t} p_t(\mathbf{x}_t)\|^2 d\mathbf{x} dt < \infty, \quad (81)$$

469 for any $R > 0$ and $0 < t_0 \leq T$;

- 470 (iv) For each $S \in (0, T)$ and all $\|\mathbf{x}_t\| \leq N_R$ and $\|\mathbf{x}'_t\| \leq N_R$, there is a constant C_{S, N_R} such
- 471 that $\nabla \log p_t(\mathbf{x}_t)$ is locally Lipschitz, i.e.,

$$\|\nabla \log p_t(\mathbf{x}_t) - \nabla \log p_t(\mathbf{x}'_t)\| \leq C_{S, N_R} \|\mathbf{x}_t - \mathbf{x}'_t\|, \quad (82)$$

472 for all $t \in (S, T)$.

473 **Remarks on Assumption 4.** Conditions (i)-(iii) are technical conditions on the forward SDE. They
 474 ensure that if we run a solution $p_t(\mathbf{x}_t)$ to the forward SDE, then $p_{T-t}(\mathbf{x}_{T-t})$ will be a solution to the
 475 reverse SDE. The last condition ensures that the solutions to the reverse SDE are unique. Assumption
 476 4 can be expected to hold in practice, i.e., for any affine $\mathbf{a}(\cdot, t)$ and bounded data manifold.

477 **Lemma 1 (Theorem 2 of Pidstrigach [2022]).** Given a forward SDE with marginals $p_t(\mathbf{x}_t)$ and an
 478 approximated score $\mathbf{s}_\theta(\mathbf{x}_t, t)$ to $\nabla \log p_t(\mathbf{x}_t)$, if the approximation error $\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla \log p_t(\mathbf{x}_t)\|$
 479 is bounded and Assumption 4 holds, then the marginal distribution of the reverse process using the
 480 approximated score starting from $p_T(\mathbf{x}_T)$ will have the same support as the data distribution $p_0(\mathbf{x}_0)$.

481 **Theorem 2.** For any given $\rho \in (0, 1)$, suppose that there exists a constant δ such that the error in the
 482 score estimation can be bounded as:

$$\|\tilde{\mathbf{s}}_\theta(\mathbf{x}_t, t, \rho) + \mathbf{G}_t - \nabla_{\mathbf{x}_t} \log \tilde{p}_t(\mathbf{x}_t | \rho)\| \leq \delta, \quad \forall \mathbf{x}_t. \quad (83)$$

483 For samples $\tilde{\mathbf{x}}_{sample} \sim p_{sample}(\mathbf{x}_0 | \rho)$ generated by the reverse process

$$d\mathbf{x}_t = [\mathbf{a}(\mathbf{x}_t, t) - b(t)^2(\tilde{\mathbf{s}}_\theta(\mathbf{x}_t, t, \rho) + \mathbf{G}_t)] dt + b(t)d\bar{\mathbf{B}}_t, \quad (84)$$

484 with prior $p_{prior} = \mathcal{N}(\mathbf{0}, \mathbf{I})$, affine drift coefficients $\mathbf{a}(\cdot, t)$, and

$$\tilde{\mathbf{s}}_\theta(\mathbf{x}_t, t, \rho) = (1 + w)\mathbf{s}_\theta(\mathbf{x}_t, t, \rho) - w\mathbf{s}_\theta(\mathbf{x}_t, t, \emptyset), \quad (85)$$

485 as $T \rightarrow \infty$, $p_{sample}(\mathbf{x}_0 | \rho)$ will have the same support as $\tilde{p}_0(\mathbf{x}_0 | \rho)$. Further, as $\beta \rightarrow \infty$, $\tilde{\mathbf{x}}_{sample}$
 486 will concentrate around $\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{D}_\rho} f(\mathbf{x})$.

487 *Proof.*

488 For the forward process $d\mathbf{x}_t = \mathbf{a}(\mathbf{x}_t, t)dt + b(t)d\mathbf{B}_t$, $t \in [0, T]$ with affine drift coefficients $\mathbf{a}(\cdot, t)$,
 489 conditions (i)-(ii) in Assumption 4 are satisfied. For the given data set $\{\mathbf{x}^{(i)}\}_{i=1}^N$ contained in a ball of
 490 radius M_R , we have that $\log \tilde{p}_t(\mathbf{x}_t, t) \in \mathcal{C}^\infty$ in both t and \mathbf{x}_t for $t > 0$ where the product distribution
 491 $\tilde{p}_0(\mathbf{x}_0 | \rho) \propto p_0(\mathbf{x}_0 | \rho)B_\beta(\mathbf{x}_0)$. Therefore we can integrate \tilde{p}_t and its derivative over compact sets,
 492 implying that condition (iii) holds. Furthermore, for each $S \in (0, T)$, the Hessian w.r.t. (\mathbf{x}_t, t) is
 493 continuous and obtains its maximum and minimum on the compact set $[S, T] \times B_{N_R}$, where B_{N_R}
 494 is the ball of diameter N_R around the origin. Therefore, the gradient $\nabla \log \tilde{p}_t(\mathbf{x}_t)$ is Lipschitz on
 495 $[S, T] \times B_{N_R}$, which proves condition (iv).

496 The stationary distribution of the forward process is characterized by the corresponding Fokker-
 497 Planck equations, where $p_T = \mathcal{N}(\mathbf{0}, \mathbf{I})$ when $T \rightarrow \infty$. Then we have that $p_{prior} = p_T$. Based
 498 on Lemma 1, if the score matching error is bounded, then the sampling distribution $p_{sample}(\mathbf{x}_0 | \rho)$
 499 with prior $p_{prior} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ will have the same support as the product distribution $\tilde{p}_0(\mathbf{x}_0 | \rho) \propto$
 500 $p_0(\mathbf{x}_0 | \rho)B_\beta(\mathbf{x}_0)$, where B_β is the Boltzmann distribution $B_\beta(\mathbf{x}_0) \propto e^{-\beta f(\mathbf{x}_0)}$.

Since $p_0(\mathbf{x}_0|\rho)$ has support \mathcal{D}_ρ and the Boltzmann factor only changes the relative density within that domain, the support of $\tilde{p}_0(\mathbf{x}_0|\rho)$ also remains \mathcal{D}_ρ , i.e.,

$$\text{supp } p_{\text{sample}}(\mathbf{x}_0|\rho) = \text{supp } \tilde{p}_0(\mathbf{x}_0|\rho) = \mathcal{D}_\rho. \quad (86)$$

As $\beta \rightarrow \infty$, sampling from the product distribution $\tilde{p}_0(\mathbf{x}_0|\rho)$ is equivalent to solving the optimization problem $\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{D}_\rho} f(\mathbf{x})$. Then we have that as $T \rightarrow \infty$ and $\beta \rightarrow \infty$, the sample $\tilde{\mathbf{x}}_{\text{sample}}$ will concentrate around \mathbf{x}^* .

□

Theorem 2 establishes that, by introducing an additional gradient guidance term into the reverse process, the sampling distribution of GGDOpt will attain the exact same support as the data distribution. Moreover, as the inverse temperature parameter β increases, the sampling distribution becomes increasingly concentrated around points with the lowest function values within the support of the data distribution.

The assumption in score estimation quantifies the approximation accuracy of the trained score network relative to the true score function. It depends on the training quality of the neural network and the expressiveness of the model class and this type of assumption is common in the theoretical analysis of diffusion models (see, e.g., Pidstrigach [2022], De Bortoli et al. [2021]) and is used to establish convergence results in generative modeling and sampling.

E.4 Proof of Theorem 3

Lemma 2 (Bolley and Villani [2005]). Let ν be a probability measure on \mathbb{R}^d . Assume that there exist \mathbf{x}_0 and a constant $\alpha > 0$ such that

$$\int e^{\alpha \|\mathbf{x} - \mathbf{x}_0\|_2^2} d\nu(\mathbf{x}) < \infty. \quad (87)$$

Then for any probability measure μ on \mathbb{R}^d , it satisfies

$$\mathcal{W}_2(\mu, \nu) \leq C_\nu (\sqrt{D_{\text{KL}}(\mu||\nu)} + (D_{\text{KL}}(\mu||\nu)/2)^{1/4}), \quad (88)$$

where \mathcal{W}_2 is the 2-Wasserstein distance and C_ν is defined as

$$C_\nu = \inf_{\mathbf{x}_0 \in \mathbb{R}^d, \alpha > 0} \sqrt{\frac{1}{\alpha} \left(\frac{3}{2} + \log \int e^{\alpha \|\mathbf{x} - \mathbf{x}_0\|_2^2} d\nu(\mathbf{x}) \right)}. \quad (89)$$

Lemma 3 (Polyanskiy and Wu [2016]). For any two probability density functions μ, ν with bounded second moments, let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a C^1 function such that

$$\|\nabla f(\mathbf{x})\|_2 \leq C_1 \|\mathbf{x}\|_2 + C_2, \forall \mathbf{x} \in \mathbb{R}^d, \quad (90)$$

for some constants $C_1, C_2 \geq 0$. Then

$$\left| \int_{\mathbb{R}^d} f(\mathbf{x}) d\mu - \int_{\mathbb{R}^d} f(\mathbf{x}) d\nu \right| \leq (C_1 \sigma + C_2) \mathcal{W}_2(\mu, \nu), \quad (91)$$

where \mathcal{W}_2 is the 2-Wasserstein distance and

$$\sigma^2 = \max \left\{ \int_{\mathbb{R}^d} \|\mathbf{x}\|_2^2 d\mu(\mathbf{x}), \int_{\mathbb{R}^d} \|\mathbf{x}\|_2^2 d\nu(\mathbf{x}) \right\}. \quad (92)$$

Lemma 4 (Polyanskiy and Wu [2016]). Let p_t be the time t -marginal of a Brownian motion with initial distribution μ_{data} . Denote by $c_i, i = 1, \dots, d$ the eigenvalues of the covariance matrix $\text{Cov}(\mu_{\text{data}})$. Let μ_{prior} be the normal distribution with mean $m_T = \mathbb{E}[\mu_{\text{data}}]$ and covariance $C_T = \text{Cov}[\mu_{\text{data}}] + T\mathbf{I}$. Then

$$D_{KL}(p_T || \mu_{\text{prior}}) \leq \frac{1}{2} \log \left(\frac{\prod_{i=1}^d (c_i + T)}{T^d} \right). \quad (93)$$

Assumption 1. We assume the following conditions hold:

- The forward process is given by $d\mathbf{x} = b(t)d\mathbf{B}_t$;
- The reverse process starts in $p_{prior} = \mathcal{N}(\mathbf{m}_T, \Sigma_T)$ where $\mathbf{m}_T = \mathbb{E}[\tilde{p}_0(\mathbf{x}_0|\rho)]$ and $\Sigma_T = \text{Cov}(\tilde{p}_0(\mathbf{x}_0|\rho)) + T \cdot \mathbf{I}$;
- The objective function $f(\mathbf{x})$ satisfies $\|\nabla_{\mathbf{x}} f(\mathbf{x})\|_2 \leq C_1 \|\mathbf{x}\|_2 + C_2$.

The first two conditions in Assumption 1 correspond to the Variance Exploding (VE) SDE in (Song et al. [2020b]) and are primarily used to characterize the discrepancy between the end distribution of the forward process and the prior distribution of the reverse process. Similar results can also be obtained for other forms of diffusion processes, e.g., Ornstein–Uhlenbeck processes. The third assumption imposes a growth bound on the gradient of the objective function. This type of regularity condition is common in the convergence analysis of stochastic optimization and sampling algorithms, particularly when studying stability and convergence under Langevin dynamics or diffusion-based methods (see, e.g., Raginsky et al. [2017]). In practice, this assumption holds for a broad class of functions, including smooth bounded functions and quadratic objectives, which frequently arise in real-world optimization problems.

Theorem 3. Under Assumption 1, denote $\sigma^{(k)}, k = 1, \dots, n$, the eigenvalues of Σ_T . For any given $\rho \in (0, 1)$, denote $N_\rho = |\mathcal{D}_\rho|$ and $\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{D}_\rho} f(\mathbf{x})$. Then for any given $T > 0$ and $\beta > 0$, the optimization error can be bounded by

$$|\mathbb{E}[f(\tilde{\mathbf{x}}_t)] - f(\mathbf{x}^*)| \leq \underbrace{C_I(\sqrt{C_T} + (C_T/2)^{1/4})}_{I_1} + \underbrace{(N_\rho - 1) \max_{\mathbf{x} \in \mathcal{D}_\rho} |f(\mathbf{x}) - f(\mathbf{x}^*)| e^{-\beta \delta_\rho}}_{I_2}, \quad (94)$$

where

$$\begin{aligned} C_I &= \inf_{\mathbf{y} \in \mathbb{R}^n, \alpha > 0} \left\{ \sqrt{\frac{1}{\alpha} \left(\frac{3}{2} + \log \int e^{\alpha \|\mathbf{x} - \mathbf{y}\|_2^2} \tilde{p}_0 d\mathbf{x} \right)} (C_1 \sigma_M + C_2) \right\}, \\ \sigma_M &= \max \left\{ \int_{\mathbb{R}^n} \|\mathbf{x}\|_2^2 \tilde{p}_0 d\mathbf{x}, \int_{\mathbb{R}^n} \|\mathbf{x}\|_2^2 p^\pi d\mathbf{x} \right\}, \\ C_T &= \frac{1}{2} \log \left(\prod_{k=1}^n (\sigma^{(k)}/T) \right), \\ \delta_\rho &= \min_{\mathbf{x} \in \mathcal{D}_\rho, f(\mathbf{x}) \neq f(\mathbf{x}^*)} |f(\mathbf{x}) - f(\mathbf{x}^*)|. \end{aligned} \quad (95)$$

Proof.

Firstly, we give the form of I_1 . By Lemma 4, we know that

$$D_{KL}(\tilde{p}_0 \| p_{sample}) \leq D_{KL}(p_T \| p_{prior}) \leq \frac{1}{2} \log \left(\prod_{k=1}^n (\sigma^{(k)}/T) \right) = C_T. \quad (96)$$

For $\tilde{p}_0(\mathbf{x}_0|\rho) \propto p_0(\mathbf{x}_0|\rho) B_\beta(\mathbf{x})$, there exist \mathbf{y} and a constant $\alpha > 0$ such that

$$\int e^{\alpha \|\mathbf{x} - \mathbf{y}\|_2^2} d\nu(\mathbf{x}) < \infty. \quad (97)$$

Then by Lemma 2, it satisfies

$$\begin{aligned} &\mathcal{W}_2(p_{sample}, \tilde{p}_0(\mathbf{x}_0|\rho)) \\ &\leq C_\nu \left(\sqrt{D_{KL}(p_{sample} \| \tilde{p}_0(\mathbf{x}_0|\rho))} + (D_{KL}(p_{sample} \| \tilde{p}_0(\mathbf{x}_0|\rho))/2)^{1/4} \right), \end{aligned} \quad (98)$$

where C_ν is defined as

$$C_\nu = \inf_{\mathbf{y} \in \mathbb{R}^d, \alpha > 0} \sqrt{\frac{1}{\alpha} \left(\frac{3}{2} + \log \int \exp(\alpha \|\mathbf{x} - \mathbf{y}\|_2^2) d\tilde{p}_0(\mathbf{x}|\rho) \right)}. \quad (99)$$

By Lemma 3, we have that

$$|\mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \mathbb{E}[f(\mathbf{x}^\pi)]| \leq C_I(\sqrt{C_T} + (C_T/2)^{1/4}). \quad (100)$$

Next, we show the form of I_2 . for $\mathbf{x}^{(i)} \in \text{supp } \tilde{p}_0(\mathbf{x}_0|\rho)$, the probability is given by

$$p^{(i)} = \frac{e^{-\beta f(\mathbf{x}^{(i)})}}{\sum_{i=1}^{N_\rho} e^{-\beta f(\mathbf{x}^{(i)})}}. \quad (101)$$

Denote $f^* = \min_{i=1}^{N_\rho} f(\mathbf{x}^{(i)})$ and $\text{Ind} = \{i \mid f(\mathbf{x}^{(i)}) = f^*\}$. Let $\delta^{(i)} = f(\mathbf{x}^{(i)}) - f^* \geq 0$, then the probability can be expressed as

$$p^{(i)} = \frac{e^{-\beta f(\mathbf{x}^{(i)})}}{\sum_{i=1}^{N_\rho} e^{-\beta f(\mathbf{x}^{(i)})}} = \frac{e^{-\beta(f^* + \delta^{(i)})}}{\sum_{i=1}^{N_\rho} e^{-\beta(f^* + \delta^{(i)})}} = \frac{e^{-\beta \delta^{(i)}}}{\sum_{i=1}^{N_\rho} e^{-\beta \delta^{(i)}}}. \quad (102)$$

Then we have

$$\mathbb{E}[f(\mathbf{x})] = \sum_{i=1}^{N_\rho} f(\mathbf{x}^{(i)}) p^{(i)} = \sum_{i=1}^{N_\rho} (f^* + \delta^{(i)}) \frac{e^{-\beta \delta^{(i)}}}{\sum_{i=1}^{N_\rho} e^{-\beta \delta^{(i)}}}, \quad (103)$$

and the limited inverse temperature error is given by

$$\begin{aligned} |\mathbb{E}[f(\mathbf{x})] - f^*| &= \left| \sum_{i=1}^{N_\rho} (f^* + \delta^{(i)}) \frac{e^{-\beta \delta^{(i)}}}{\sum_{i=1}^{N_\rho} e^{-\beta \delta^{(i)}}} - \sum_{i=1}^{N_\rho} f^* \frac{e^{-\beta \delta^{(i)}}}{\sum_{i=1}^{N_\rho} e^{-\beta \delta^{(i)}}} \right| \\ &= \sum_{i=1}^{N_\rho} \delta^{(i)} \frac{e^{-\beta \delta^{(i)}}}{\sum_{i=1}^{N_\rho} e^{-\beta \delta^{(i)}}}. \end{aligned} \quad (104)$$

Note that $\delta^{(i)} = 0$ for $i \in \text{Ind}$, so we can simplify the sum as

$$\sum_{i=1}^{N_\rho} \delta^{(i)} \frac{e^{-\beta \delta^{(i)}}}{\sum_{i=1}^{N_\rho} e^{-\beta \delta^{(i)}}} = \sum_{i=1, i \notin \text{Ind}}^{N_\rho} \delta^{(i)} \frac{e^{-\beta \delta^{(i)}}}{\sum_{i=1, i \notin \text{Ind}}^{N_\rho} e^{-\beta \delta^{(i)}} + \sum_{i=1, i \in \text{Ind}}^{N_\rho} e^{-\beta \delta^{(i)}}}. \quad (105)$$

The denominator

$$\sum_{i=1, i \notin \text{Ind}}^{N_\rho} e^{-\beta \delta^{(i)}} + \sum_{i=1, i \in \text{Ind}}^{N_\rho} e^{-\beta \delta^{(i)}} = \sum_{i=1, i \notin \text{Ind}}^{N_\rho} e^{-\beta \delta^{(i)}} + |\text{Ind}| \geq 1, \quad (106)$$

so we have that

$$\begin{aligned} |\mathbb{E}[f(\mathbf{x})] - f^*| &= \sum_{i=1}^{N_\rho} \delta^{(i)} \frac{e^{-\beta \delta^{(i)}}}{\sum_{i=1}^{N_\rho} e^{-\beta \delta^{(i)}}} \\ &\leq \sum_{i=1, i \notin \text{Ind}}^{N_\rho} \delta^{(i)} e^{-\beta \delta^{(i)}} \\ &\leq (N_\rho - 1) \max_{\mathbf{x} \in \mathcal{D}_\rho} |f(\mathbf{x}) - f(\mathbf{x}^*)| e^{-\beta \delta_\rho}, \end{aligned} \quad (107)$$

where

$$\delta_\rho = \min_{\mathbf{x} \in \mathcal{D}_\rho, f(\mathbf{x}) \neq f(\mathbf{x}^*)} |f(\mathbf{x}) - f(\mathbf{x}^*)|. \quad (108)$$

Then the optimization error can be bounded by

$$|\mathbb{E}[f(\tilde{\mathbf{x}}_t)] - f(\mathbf{x}^*)| \leq \underbrace{C_I (\sqrt{C_T} + (C_T/2)^{1/4})}_{I_1} + \underbrace{(N_\rho - 1) \max_{\mathbf{x} \in \mathcal{D}_\rho} |f(\mathbf{x}) - f(\mathbf{x}^*)| e^{-\beta \delta_\rho}}_{I_2}, \quad (109)$$

where

$$\begin{aligned} C_I &= \inf_{\mathbf{y} \in \mathbb{R}^n, \alpha > 0} \left\{ \sqrt{\frac{1}{\alpha} \left(\frac{3}{2} + \log \int e^{\alpha \|\mathbf{x} - \mathbf{y}\|_2^2} \tilde{p}_0 d\mathbf{x} \right)} (C_1 \sigma_M + C_2) \right\}, \\ \sigma_M &= \max \left\{ \int_{\mathbb{R}^n} \|\mathbf{x}\|_2^2 \tilde{p}_0 d\mathbf{x}, \int_{\mathbb{R}^n} \|\mathbf{x}\|_2^2 p^\pi d\mathbf{x} \right\}, \\ C_T &= \frac{1}{2} \log \left(\prod_{k=1}^n (\sigma^{(k)}/T) \right), \\ \delta_\rho &= \min_{\mathbf{x} \in \mathcal{D}_\rho, f(\mathbf{x}) \neq f(\mathbf{x}^*)} |f(\mathbf{x}) - f(\mathbf{x}^*)|. \end{aligned} \quad (110)$$

566

□

567 Theorem 3 establishes that, in practical settings, the optimization error of the sampling process can
 568 be decomposed and bounded by two components: the limited time length error I_1 and the limited
 569 inverse temperature error I_2 , which are given as follows:

$$|\mathbb{E}[f(\tilde{\mathbf{x}}_t)] - f(\mathbf{x}^*)| \leq \underbrace{|\mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \mathbb{E}[f(\mathbf{x}^\pi)]|}_{I_1} + \underbrace{|\mathbb{E}[f(\mathbf{x}^\pi)] - f(\mathbf{x}^*)|}_{I_2}. \quad (111)$$

570 As a direct corollary, under mild assumptions, GGDOpt is shown to generate asymptotically optimal
 571 solutions to problem (17) as the time length T and inverse temperature β increase.

References

- Xiaodi Bai, Jie Sun, and Xiaojin Zheng. An augmented lagrangian decomposition method for chance-constrained optimization problems. *INFORMS Journal on Computing*, 33(3):1056–1069, 2021.
- Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88:411–424, 2000.
- François Bolley and Cédric Villani. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pages 331–352, 2005.
- Pafnutii Lvovich Chebyshev. Des valeurs moyennes. *J. Math. Pures Appl*, 12(2):177–184, 1867.
- Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022.
- Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.
- Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.
- Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2008.
- Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *arXiv preprint arXiv:2404.14743*, 2024.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- B Klartag and S Sodin. Variations on the berry–esseen theorem. *Theory of Probability & Its Applications*, 56(3):403–419, 2012.
- Siddarth Krishnamoorthy, Satvik Mehul Mashkaria, and Aditya Grover. Diffusion models for black-box optimization. In *International Conference on Machine Learning*, pages 17842–17857. PMLR, 2023.
- Yang Li, Jinpei Guo, Runzhong Wang, Hongyuan Zha, and Junchi Yan. Fast t2t: Optimization consistency speeds up diffusion-based training-to-testing solving for combinatorial optimization. *Advances in Neural Information Processing Systems*, 37:30179–30206, 2024.
- Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.
- Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.
- Bernardo K Pagnoncelli, Shabbir Ahmed, and Alexander Shapiro. Sample average approximation method for chance constrained programming: theory and applications. *Journal of Optimization Theory and Applications*, 142(2):399–416, 2009.
- Jakiw Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

617 Yury Polyanskiy and Yihong Wu. Wasserstein continuity of entropy and outer bounds for interference
618 channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

619 Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic
620 gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages
621 1674–1703. PMLR, 2017.

622 Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical
623 image segmentation. In *Medical Image Computing and Computer-assisted Intervention–MICCAI*
624 *2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III*
625 *18*, pages 234–241. Springer, 2015.

626 Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv*
627 *preprint arXiv:2010.02502*, 2020a.

628 Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben
629 Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint*
630 *arXiv:2011.13456*, 2020b.

631 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz
632 Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing*
633 *Systems*, 30, 2017.

634 Kun-Yu Wang, Anthony Man-Cho So, Tsung-Hui Chang, Wing-Kin Ma, and Chong-Yung Chi.
635 Outage constrained robust transmit optimization for multiuser mimo downlinks: Tractable approx-
636 imations by conic optimization. *IEEE Transactions on Signal Processing*, 62(21):5690–5705,
637 2014.

638 Peng Wang, Rujun Jiang, Qingyuan Kong, and Laura Balzano. A proximal dc algorithm for sample
639 average approximation of chance constrained programming. *arXiv preprint arXiv:2301.00423*,
640 2023.

641 Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on*
642 *computer vision (ECCV)*, pages 3–19, 2018.

643 Shenglong Zhou, Lili Pan, Naihua Xiu, and Geoffrey Ye Li. A 0/1 constrained optimization solving
644 sample average approximation for chance constrained programming. *Mathematics of Operations*
645 *Research*, 2024.