CS395T: Continuous Algorithms, Part XVIII
Restricted Gaussian dynamics

Kevin Tian

1 Restricted Gaussian oracle

The central object of this lecture is the following type of oracle.

Definition 1 (Restricted Gaussian oracle)**.** *We say $\mathcal{O}$ is a* restricted Gaussian oracle *(RGO) for $\psi : \mathbb{R}^d \to \mathbb{R}$ if for any $\mathbf{v} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}_{\geq 0}$, $\mathcal{O}(\lambda, \mathbf{v})$ returns a sample from the density on $\mathbb{R}^d$*

$$\propto \exp\left(-\frac{\lambda}{2}\left\|\mathbf{x}-\mathbf{v}\right\|_2^2 - \psi(\mathbf{x})\right). \tag{1}$$

The first thing to note is the similarity between Definition 1 and that of a *proximal oracle* (Definition 7, Part II), which asks to minimize the potential $\psi + \frac{\lambda}{2}\left\|\cdot - \mathbf{v}\right\|_2^2$ rather than sample from a density defined by it. Similarly to the role of the proximal oracle in the *proximal point method* (Section 2, Part III), we will see in Section 1 of this lecture how to design samplers when granted access to an appropriate RGO. This lets us effectively decouple sampling algorithms into an "outer loop" proximal point method and an "inner loop" efficient implementation of each RGO.

In this introductory section, we first answer the question: when should we expect ψ to support an efficient RGO? Observe that this oracle effectively asks us to sample the density $\exp(-\psi(\mathbf{x}))$ "restricted" by the Gaussian $\exp(\mathbf{v}, \frac{1}{\lambda}\mathbf{I}_d)$. A basic setting is when ψ is "simple" enough, e.g., coordinatewise separable ($\psi(\mathbf{x}) = \left\|\mathbf{x}\right\|_1$ or the indicator of an axis-aligned box), or a sufficiently uncomplicated function of $\left\|\cdot\right\|_2$ that we can exactly integrate. It is relatively reasonable to assume that basic numerical operations in one dimension, such as integration, can be performed in constant time. For these types of ψ, our assumption gives an exact $O(d)$-time RGO.

The real power of the RGO framework is as a generic reduction for structured sampler design. Assuming that ψ has some additional structure, we can often use the additional regularization afforded by an RGO (via tuning λ) to attain better parameter tradeoffs for algorithms. For example, in Section 1.2 we explain how proximal sampling methods generically improve dependences on the condition number κ to linear, for high-accuracy well-conditioned samplers.

1.1 Well-conditioned densities

In this section, we prove that *extremely well-conditioned* $\psi + \frac{\lambda}{2}\left\|\cdot - \mathbf{v}\right\|_2^2$ admit efficient RGOs. For simplicity, we assume the ability to minimize the potential $\psi + \frac{\lambda}{2}\left\|\cdot - \mathbf{v}\right\|_2^2$; an efficient optimization algorithm in the well-conditioned regime we consider is provided in Theorem 4, Part II.

Our strategy is based on rejection sampling, for which we first provide a generic analysis.

Lemma 1. *Let π, μ be probability densities over the same sample space Ω, and suppose $\pi \propto P$, $\mu \propto Q$ for nonnegative functions $P : \Omega \to \mathbb{R}_{\geq 0}$, $Q : \Omega \to \mathbb{R}_{\geq 0}$. Suppose that*

$$P(\omega) \leq Q(\omega) \text{ for all } \omega \in \Omega.$$

Let $N_P := \int_\Omega P(\omega)\mathrm{d}\omega$, $N_Q := \int_\Omega Q(\omega)\mathrm{d}\omega$. There is an algorithm that samples from π, using an expected $\frac{N_Q}{N_P}$ samples from μ, $\frac{N_Q}{N_P}$ evaluations of Q, and $\frac{N_Q}{N_P}$ evaluations of P.

Proof. The algorithm is to repeatedly sample $\omega \sim \mu$, and accept the sample with probability $\frac{P(\omega)}{Q(\omega)} \le 1$. The probability that any given loop of the algorithm returns a sample is

$$\int_\Omega \mu(\omega) \cdot \frac{P(\omega)}{Q(\omega)} \mathrm{d}\omega = \int_\Omega \frac{Q(\omega)}{N_Q} \cdot \frac{P(\omega)}{Q(\omega)} \mathrm{d}\omega = \frac{N_P}{N_Q},$$

and conditioned on returning, the output density is $\propto Q(\omega) \cdot \frac{P(\omega)}{Q(\omega)} = P(\omega)$ as desired. □

We now give an RGO when the condition number of $\psi + \frac{\lambda}{2} \|\cdot - \mathbf{v}\|_2^2$ is nearly 1.[1]

Lemma 2. *Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex, and let $(\lambda, \mathbf{v}) \in \mathbb{R}_{\ge 0} \times \mathbb{R}^d$. Assume we know $\mathbf{x}^\star := \mathrm{argmin}_{\mathbf{x} \in \mathbb{R}^d} \psi(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{v}\|_2^2$. Then we can sample from the density (1) in*

$$O\left(\left(\frac{L+\lambda}{\mu+\lambda}\right)^{\frac{d}{2}}\right) \text{ expected value queries to } \psi,$$

$$\text{and } O\left(\left(\frac{L+\lambda}{\mu+\lambda}\right)^{\frac{d}{2}} d\right) \text{ expected additional time.}$$

Proof. Throughout this proof let $U(\mathbf{x}) := \psi(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{v}\|_2^2$. We instantiate Lemma 1 with

$$P(\mathbf{x}) \leftarrow \exp\left(-U(\mathbf{x})\right),\ Q(\mathbf{x}) \leftarrow \exp\left(-U(\mathbf{x}^\star) - \frac{\mu+\lambda}{2} \|\mathbf{x} - \mathbf{x}^\star\|_2^2\right).$$

Under our computational model, we can sample exactly from $\mu \propto Q$ in $O(d)$ time, because $\mu = \mathcal{N}(\mathbf{x}^\star, \frac{1}{\mu+\lambda})$. The fact that $P \le Q$ pointwise follows from strong convexity. Finally, following notation from Lemma 1, we can bound

$$\frac{N_Q}{N_P} \le \frac{\int \exp\left(-U(\mathbf{x}^\star) - \frac{\mu+\lambda}{2} \|\mathbf{x} - \mathbf{x}^\star\|_2^2\right) \mathrm{d}\mathbf{x}}{\int \exp\left(-U(\mathbf{x}^\star) - \frac{L+\lambda}{2} \|\mathbf{x} - \mathbf{x}^\star\|_2^2\right) \mathrm{d}\mathbf{x}} = \left(\frac{L+\lambda}{\mu+\lambda}\right)^{\frac{d}{2}},$$

where the inequality above used the pointwise bound $U(\mathbf{x}) \le U(\mathbf{x}^\star) + \frac{L+\lambda}{2} \|\mathbf{x} - \mathbf{x}^\star\|_2^2$, and the equality plugged in the exact normalizing constants for multivariate Gaussians. □

Lemma 2 shows that when ψ is L-smooth and μ-strongly convex, we can implement an RGO for it using only a *constant* number of calls to a value oracle, in regimes where the regularization parameter $\lambda \gg Ld$ is sufficiently large. Indeed, in this regime, $\frac{L+\lambda}{\mu+\lambda} = 1 + O(\frac{1}{d})$, and hence the expected value query complexity in Lemma 2 is $O(1)$. This theme of using the additional regularization afforded by an RGO to enable more efficient algorithms is prominent in this framework. Of course, there is a tradeoff: larger λ means the density (1) drifts further from that $\propto \exp(-\psi)$, requiring more outer loop iterations. This point will be discussed in depth in Section 2.

One other point we remark on is that Lemma 2 only yields an *inexact* proximal point oracle. This is because rejection sampling is never guaranteed to return in finite time (though, if it is expected to return in T iterations, then it will with probability $\ge 1 - \delta$ in $O(T \log(\frac{1}{\delta}))$ iterations by standard binomial concentration). This typically does not pose an issue when RGOs are used in an outer loop with total variation distance guarantees. In particular, any total variation error in the RGO implementation can be "charged" to the overall error via a union bound argument (see discussion after Fact 1, Part XIV). When the outer convergence metric is more complex, e.g., a Rényi divergence, more careful arguments can sometimes still be used to retain these alternative guarantees under inexact RGO implementations (cf. Lemma 5.2, [AC24]).

[1] The RGO implementation in Lemma 2 is a simplification of the original construction in Section 4, [LST21], due to Nima Anari as developed in his course notes [Ana23].

1.2 Generic reduction framework

To give the reader some sense of how to apply the result of Lemma 2 to design efficient sampling algorithms, let us first state a result we will prove in Section 2.

Theorem 1 (Well-conditioned proximal sampling)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex, let $\pi^\star \propto \exp(-V)$, and let $\eta \le \frac{1}{\mu}$. Consider the algorithm which initializes $\mathbf{x}^{(0)} \leftarrow \text{argmin}_{\mathbf{x}\in\mathbb{R}^d} V(\mathbf{x})$, and then iterates for all $0 \le k < K$:*

$$\begin{aligned} \mathbf{y}^{(k)} &\sim \textit{ density } \propto \exp\left(-\frac{1}{2\eta}\left\|\mathbf{x}^{(k)} - \cdot\right\|_2^2\right), \\ \mathbf{x}^{(k+1)} &\sim \textit{ density } \propto \exp\left(-V(\cdot) - \frac{1}{2\eta}\left\|\mathbf{y}^{(k)} - \cdot\right\|_2^2\right). \end{aligned} \tag{2}$$

Then letting $\pi^{(K)}$ denote the law of $\mathbf{x}^{(K)}$,

$$D_{\text{KL}}\left(\pi^{(K)} \| \pi^\star\right) \le \epsilon^2, \textit{ for } K \ge \frac{2}{\eta\mu}\log\left(\frac{d}{\eta\mu\epsilon^2}\right).$$

To give some intuition for the updates in (2), observe that they alternate sampling from the conditional marginals of the density on $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d$ that is

$$\propto \exp\left(-V(\mathbf{x}) - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|_2^2\right).$$

This is the joint density of $\mathbf{x} \sim \pi^\star$ and $\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\mathbf{x}, \eta\mathbf{I}_d)$, so the $\mathbf{x}$ marginal is precisely $\pi^\star$. In Section 2, we show that Theorem 1 implements a reversible Markov chain with stationary measure $\pi^\star$, and explicitly provide a bound on its rate of relative entropy decay in each iteration.

Each $\mathbf{y}^{(k)}$ iterate in Theorem 1 is straightforward to sample: it is just a scaled Gaussian centered at $\mathbf{x}^{(k)}$. On the other hand, updating to $\mathbf{x}^{(k+1)}$ is effectively implementing an RGO with $\psi \leftarrow V$, $\mathbf{v} \leftarrow \mathbf{y}^{(k)}$, and $\lambda \leftarrow \frac{1}{\eta}$. Notably, the presence of the quadratic $\frac{1}{2\eta}\left\|\cdot - \mathbf{y}^{(k)}\right\|_2^2$ can significantly improve the conditioning of this RGO subproblem for a judicious choice of η.

Let us give a sense of how to apply Theorem 1, starting with the well-conditioned setting.

Corollary 1. *Let $V : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex, and let $\pi^\star \propto \exp(-V)$, $\kappa := \frac{L}{\mu}$, $\epsilon \in (0, 1)$. There is an algorithm that samples within ϵ total variation distance of $\pi^\star$, using*

$$O\left(\kappa d \log^2\left(\frac{\kappa d}{\epsilon}\right)\right) \textit{ queries to } V.$$

Proof. The algorithm is Theorem 1, with $\eta \leftarrow \frac{1}{Ld}$, and with final KL divergence error $\epsilon^2 \leftarrow \frac{1}{2}\epsilon^2$. By Pinsker's inequality, this implies the final distribution $\pi^{(K)}$ has total variation distance $\le \frac{\epsilon}{2}$ from $\pi^\star$. We implement each RGO required by Theorem 1 with failure probability $\frac{\epsilon}{2K}$ by boosting the success probability of Lemma 2. Because the condition number

$$\frac{L + \frac{1}{\eta}}{\mu + \frac{1}{\eta}} \le \frac{L(d+1)}{Ld} \le 1 + \frac{1}{d}$$

is bounded, each RGO call only requires $O(\log(\frac{\kappa d}{\epsilon}))$ queries to achieve $\frac{\epsilon}{2K}$ failure probability. □

Corollary 1 gives a simple recipe for quadratically improving our somewhat tedious analysis of MALA from Section 3, Part XVII (which admittedly was quite loose). It replaced our challenging task, sampling from $\pi^\star$, to sampling from $\approx \kappa d$ regularized densities, each so well-conditioned that we could obtain a sample (via Lemma 2) using nearly a constant number of queries.

More interestingly, Corollary 1 can be further improved if an algorithm is designed with improved dependence on d (regardless of its native dependence on κ). This is done via a simple reduction.

Corollary 2. *If there is an algorithm which, given oracle query access to L-smooth and μ-strongly convex $V : \mathbb{R}^d \to \mathbb{R}$, uses a total of $\mathcal{T}(\kappa, d, \epsilon)$ queries to achieve total variation distance ϵ, there is an algorithm using $K \cdot \mathcal{T}(2, d, \frac{\epsilon}{2K})$ queries to achieve total variation distance ϵ, for $K = O(\kappa\log(\frac{\kappa d}{\epsilon}))$.*

Proof. We apply Theorem 1 with the stated value of K, and $\eta = \frac{1}{L}$, such that it guarantees $\frac{\epsilon}{2}$ total variation distance from $\pi^\star$ assuming exact RGOs. For this value of η, all subproblems in (2) are either an exact Gaussian sample, or an RGO with condition number

$$\frac{L + \frac{1}{\eta}}{\mu + \frac{1}{\eta}} \le \frac{2L}{L} = 2.$$

Each of these K RGO subproblems can be implemented up to $\frac{\epsilon}{2K}$ total variation distance using $\mathcal{T}(2, d, \frac{\epsilon}{2K})$ oracle queries by assumption, and the conclusion follows from a union bound. □

For example, suppose there is an algorithm that implements sampling of κ-well-conditioned densities using $f(\kappa)\sqrt{d}$ polylog$(\frac{d}{\epsilon})$ queries to a gradient oracle, for an arbitrary function f (e.g., polynomial, exponential, or even larger). Then just applying Corollary 2 improves the runtime to $O(\kappa\sqrt{d}$ polylog$(\frac{\kappa d}{\epsilon}))$, i.e., the function $f(\kappa)$ can be reduced to linear in κ without loss of generality! In fact, the currently-fastest algorithm for sampling κ-well-conditioned densities in $\mathbb{R}^d$ is obtained in this way ([AC24], see Remark 1 and Section 4, Part XVII).

We discuss additional applications of the RGO to sampling from other structured density families, including log-Lipschitz densities, composite densities, finite sums, and general logconcave sampling, as well as proposed non-Euclidean generalizations, in Section 3.

2 Proximal point method for sampling

In this section, we prove Theorem 1. We mention that the key will to be establish a *relative entropy contraction* bound for the iteration (2), at a rate of $1 - O(\eta\mu)$ per iteration.

It is straightforward to prove that (2) mixes rapidly in χ^2 divergence at the rate of $1 - O(\eta\mu)$. Indeed, $\pi^\star$ satisfies $\Omega(\sqrt{\mu})$-isoperimetry by using Lemma 6, Part XVII, and $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ with $\|\mathbf{x} - \mathbf{x}'\|_2 = \Theta(\sqrt{\eta})$ have transition distributions that overlap by $\Omega(1)$ in total variation, using a similar calculation to Eq. (18), Part XVII.[2] This yields a $\Omega(\sqrt{\eta\mu})$ conductance bound (Proposition 4, Part XV), a.k.a. an $\Omega(\eta\mu)$ contraction in relative variance (Proposition 1, Part XV).

It is more subtle to prove that (2) mixes rapidly in relative entropy. We will in fact show that these updates contract in W_2^2, which due to the specific form of the iterations, we can then relate to D_{KL} error. As a starting point, we show that (2) is reversible with respect to $\pi^\star$.

Lemma 3. *In the setting of Theorem 1, the updates (2) (viewed as a transition distribution from $\mathbf{x}^{(k)} \to \mathbf{x}^{(k+1)}$) are reversible with respect to $\pi^\star$ (and hence have stationary density $\pi^\star$).*

Proof. Let $\pi^\star_{\text{ext}}$ be the density on $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d$ that is

$$\propto \exp\left(-V(\mathbf{x}) - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|_2^2\right).$$

Observe that from $\mathbf{x}^{(k)} \leftarrow \mathbf{x}$, the product of $\pi^\star(\mathbf{x})$ and $\mathcal{T}_{\mathbf{x}}(\mathbf{x}')$ for any $\mathbf{x}' \in \mathbb{R}^d$ is

$$\int \pi^\star(\mathbf{x})\pi^\star_{\text{ext}}((\mathbf{x}, \mathbf{y}) \mid \mathbf{x})\pi^\star_{\text{ext}}((\mathbf{x}', \mathbf{y}) \mid \mathbf{y})\mathrm{d}\mathbf{y} = \int \frac{\pi^\star_{\text{ext}}(\mathbf{x}, \mathbf{y})\pi^\star_{\text{ext}}(\mathbf{x}', \mathbf{y})}{\int_{\mathbf{z}} \pi^\star_{\text{ext}}(\mathbf{z}, \mathbf{y})\mathrm{d}\mathbf{z}}\mathrm{d}\mathbf{x}.$$

This is because the transition density is one round of alternating sampling from the conditional marginals of $\pi^\star_{\text{ext}}$. The final expression is a symmetric function of $\mathbf{x}, \mathbf{x}'$, and hence the Markov chain is reversible with respect to $\pi^\star$. Stationarity follows from Eq. (4), Part XV. □

Next, we provide the claimed W_2^2 contraction bound.

Lemma 4. *In the setting of Theorem 1, let $\pi^{(k)}$ denote the law of $\mathbf{x}^{(k)}$ for all $0 \le k \le K$. Then,*

$$W_2^2\left(\pi^{(k+1)}, \pi^\star\right) \le \frac{1}{(1 + \eta\mu)^2} W_2^2\left(\pi^{(k)}, \pi^\star\right) \text{ for all } 0 \le k < K.$$

[2] Formally, the $\mathbf{y}_k$ step in (2) has this amount of overlap for nearby $\mathbf{x}_k$, but it is straightforward to show that the $\mathbf{x}_{k+1}$ step cannot worsen the overlap, e.g., via Lemma 2, Part XV.

Proof. To ease notation throughout this proof, for all $\mathbf{y} \in \mathbb{R}^d$ we define the induced density

$$\pi_y^\star(\mathbf{x}) \propto \exp\left(-V(\mathbf{x}) - \frac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|_2^2\right). \tag{3}$$

We will construct a coupling between the iterates $\mathbf{x}^{(k+1)}$, obtained via running (2) from $\mathbf{x}^{(k)} \sim \pi^{(k)}$, and $\mathbf{x}_\star^{(k+1)}$, obtained via running (2) from $\mathbf{x}^{(k)} \leftarrow \mathbf{x}_\star^{(k)} \sim \pi^\star$, that witnesses the claimed inequality. In particular, observe that $\mathbf{x}_\star^{(k+1)} \sim \pi^\star$, as $\pi^\star$ is the stationary density for (2) (Lemma 3). We also let $\mathbf{y}_\star^{(k)}$ denote the intermediate iterate in (2) initialized at $\mathbf{x}_\star^{(k)}$.

First, let $\Gamma^{(k)}$ be the optimal coupling between $\pi^{(k)}$ and $\pi^\star$ that achieves $W_2^2(\pi^{(k)}, \pi^\star)$, i.e.,

$$\mathbb{E}_{(\mathbf{x}^{(k)}, \mathbf{x}_\star^{(k)}) \sim \Gamma^{(k)}}\left[\left\|\mathbf{x}^{(k)} - \mathbf{x}_\star^{(k)}\right\|_2^2\right] = W_2^2\left(\pi^{(k)}, \pi^\star\right).$$

By coupling the random Gaussian $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ used in the updates $\mathbf{y}^{(k)} \leftarrow \mathbf{x}^{(k)} + \sqrt{\eta}\boldsymbol{\xi}$ and $\mathbf{y}_\star^{(k)} \leftarrow \mathbf{x}_\star^{(k)} + \sqrt{\eta}\boldsymbol{\xi}$ in the first line of (2), we obtain a coupling $\Gamma^{(k+\frac{1}{2})}$ of $(\mathbf{y}^{(k)}, \mathbf{y}_\star^{(k)})$ satisfying

$$\mathbb{E}_{(\mathbf{y}^{(k)}, \mathbf{y}_\star^{(k)}) \sim \Gamma^{(k+\frac{1}{2})}}\left[\left\|\mathbf{y}^{(k)} - \mathbf{y}_\star^{(k)}\right\|_2^2\right] = W_2^2\left(\pi^{(k)}, \pi^\star\right).$$

Next, $\pi^\star_{\mathbf{y}^{(k)}}$ and $\pi^\star_{\mathbf{y}_\star^{(k)}}$ defined via (3) are both $\mu + \frac{1}{\eta}$-strongly logconcave. Thus, they satisfy Talagrand's transportation inequality (Lemma 13, Part XVII) and a log-Sobolev inequality (Theorem 4, Part XVII). This yields the comparison, for any pair of $(\mathbf{y}^{(k)}, \mathbf{y}_\star^{(k)}) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\begin{aligned}
W_2^2\left(\pi^\star_{\mathbf{y}^{(k)}}, \pi^\star_{\mathbf{y}_\star^{(k)}}\right) &\le \frac{2}{\mu + \frac{1}{\eta}} D_{\mathrm{KL}}\left(\pi^\star_{\mathbf{y}^{(k)}} \| \pi^\star_{\mathbf{y}_\star^{(k)}}\right) \\
&\le \frac{1}{\left(\mu + \frac{1}{\eta}\right)^2} \mathbb{E}_{\mathbf{x} \sim \pi^\star_{\mathbf{y}^{(k)}}}\left[\left\|\nabla \log\left(\frac{\pi^\star_{\mathbf{y}^{(k)}}(\mathbf{x})}{\pi^\star_{\mathbf{y}_\star^{(k)}}(\mathbf{x})}\right)\right\|_2^2\right] \\
&= \frac{1}{\left(\mu + \frac{1}{\eta}\right)^2} \mathbb{E}_{\mathbf{x} \sim \pi^\star_{\mathbf{y}^{(k)}}}\left[\left\|\nabla\left(\frac{1}{2\eta}\left\|\mathbf{x} - \mathbf{y}^{(k)}\right\|_2^2 - \frac{1}{2\eta}\left\|\mathbf{x} - \mathbf{y}^{(k)}\right\|\right)\right\|_2^2\right] \\
&= \frac{1}{(1+\eta\mu)^2}\left\|\mathbf{y}^{(k)} - \mathbf{y}_\star^{(k)}\right\|_2^2.
\end{aligned} \tag{4}$$

The first line used Talagrand's transportation inequality, the second used the log-Sobolev inequality (in Lemma 8, Part XVI), the third used the formula for $\pi^\star_{\mathbf{y}}$ in (3) (as normalizing constants do not affect $\nabla \log$), and the fourth used that the gradient is independent of $\mathbf{x} \sim \pi^\star_{\mathbf{y}^{(k)}}$.

Now, consider the coupling that draws $(\mathbf{y}^{(k)}, \mathbf{y}_\star^{(k)}) \sim \Gamma^{(k+\frac{1}{2})}$ and then conditionally couples $\mathbf{x}^{(k+1)} \sim \pi^\star_{\mathbf{y}^{(k)}}$ and $\mathbf{x}_\star^{(k+1)} \sim \pi^\star_{\mathbf{y}_\star^{(k)}}$ in the way that achieves (4). This is an overall coupling of $\pi^{(k+1)}$ and $\pi^\star$ that witnesses the contraction in the lemma statement. □

We additionally require that the KL divergence satisfies the *data processing inequality.* Intuitively, this well-known property says that any joint postprocessing of random variables can only improve the KL divergence between them. The proof is a simple application of Jensen's inequality.

Lemma 5. *Let Ω be some sample space that distributions P, Q are supported on. Let $f : \Omega \to \Omega'$ be arbitrary, and let P^f denote the density of $f(\omega)$ for $\omega \sim P$, and similarly define Q^f. Then,*

$$D_{\mathrm{KL}}\left(P^f \| Q^f\right) \le D_{\mathrm{KL}}\left(P \| Q\right).$$

Proof. Let $\Omega_{\omega'} := \{\omega \in \Omega \mid f(\omega) = \omega'\}$ be the preimage of $\omega' \in \Omega'$. Then,

$$\begin{aligned}
D_{\mathrm{KL}}\left(P^f \| Q^f\right) &= \int_{\Omega'} \log\left(\frac{P^f(\omega')}{Q^f(\omega')}\right) P^f(\omega')\mathrm{d}\omega' \\
&= \int_{\Omega'} \log\left(\frac{\int_{\Omega_{\omega'}} P(\omega)\mathrm{d}\omega}{\int_{\Omega_{\omega'}} Q(\omega)\mathrm{d}\omega}\right)\left(\int_{\Omega_{\omega'}} P(\omega)\mathrm{d}\omega\right)\mathrm{d}\omega' \\
&\le \int_{\Omega'}\left(\int_{\Omega_{\omega'}} \log\left(\frac{P(\omega)}{Q(\omega)}\right)P(\omega)\mathrm{d}\omega\right)\mathrm{d}\omega' = D_{\mathrm{KL}}\left(P\|Q\right).
\end{aligned}$$

The only inequality viewed $\log(\frac{P(\omega)}{Q(\omega)})P(\omega)$ as a realization of $X\log(X)$ for a random variable $X = \frac{P(\omega)}{Q(\omega)}$, for ω drawn from the conditional distribution of $Q(\cdot \mid \omega \in \Omega_{\omega'})$. By convexity of the function $X \to X\log(X)$, this shows that the last line above holds:

$$\begin{aligned}
\frac{\int_{\Omega_{\omega'}} P(\omega)\mathrm{d}\omega}{\int_{\Omega_{\omega'}} Q(\omega)\mathrm{d}\omega}\log\left(\frac{\int_{\Omega_{\omega'}} P(\omega)\mathrm{d}\omega}{\int_{\Omega_{\omega'}} Q(\omega)\mathrm{d}\omega}\right) &= \mathbb{E}_{\omega\sim Q(\cdot|\Omega_{\omega'})}\left[\frac{P(\omega)}{Q(\omega)}\right]\log\left(\mathbb{E}_{\omega\sim Q(\cdot|\Omega_{\omega'})}\left[\frac{P(\omega)}{Q(\omega)}\right]\right) \\
&\le \mathbb{E}_{\omega\sim Q(\cdot|\Omega_{\omega'})}\left[\frac{P(\omega)}{Q(\omega)}\log\left(\frac{P(\omega)}{Q(\omega)}\right)\right] \\
&= \frac{1}{\int_{\Omega_{\omega'}} Q(\omega)\mathrm{d}\omega}\cdot\int_{\Omega_{\omega'}}\log\left(\frac{P(\omega)}{Q(\omega)}\right)P(\omega)\mathrm{d}\omega.
\end{aligned}$$

□

This yields the following corollary on KL contraction for Markov processes.

Lemma 6. *Let P_ξ, Q_ξ be distributions supported on Ω, indexed by a random variable $\xi \sim \pi$. Let $\widetilde{P}$ be the joint distribution of (ω, ξ) for $\xi \sim \pi$ and then $\omega \sim P_\xi$, and similarly define $\widetilde{Q}$. Finally, let P, Q be the marginal distributions of $\widetilde{P}, \widetilde{Q}$ on Ω (i.e., averaged over $\xi \sim \pi$). Then,*

$$D_{\mathrm{KL}}\left(P\|Q\right) \le \mathbb{E}_{\xi\sim\pi}\left[D_{\mathrm{KL}}\left(P_\xi\|Q_\xi\right)\right].$$

Proof. This follows from

$$\begin{aligned}
D_{\mathrm{KL}}\left(P\|Q\right) \le D_{\mathrm{KL}}\left(\widetilde{P}\|\widetilde{Q}\right) &= \mathbb{E}_{\xi\sim\pi}\left[\mathbb{E}_{\omega\sim P_\xi}\left[\log\left(\frac{\widetilde{P}(\omega,\xi)}{\widetilde{Q}(\omega,\xi)}\right)\right]\right] \\
&= \mathbb{E}_{\xi\sim\pi}\left[\mathbb{E}_{\omega\sim P_\xi}\left[\log\left(\frac{P_\xi(\omega)}{Q_\xi(\omega)}\right)\right]\right] = \mathbb{E}_{\xi\sim\pi}\left[D_{\mathrm{KL}}\left(P_\xi\|Q_\xi\right)\right].
\end{aligned}$$

where we used the data processing inequality (Lemma 5) on the function $f(\omega, \xi) = \omega$. □

We are now in good shape to complete our proof of Theorem 1.

Proof of Theorem 1. First, recall from Lemma 5, Part XVI that $W_2^2(\pi^{(0)}, \pi^\star) \le \frac{2d}{\mu}$ for our choice of initialization (i.e., a point mass at the minimizer of V). Thus, after the stated number of iterations K, we have by repeatedly applying Lemma 4 that

$$W_2^2\left(\pi^{(K-1)}, \pi^\star\right) \le W_2^2\left(\pi^{(0)}, \pi^\star\right)\cdot\frac{\epsilon^2\mu\eta}{d} \le 2\eta\epsilon^2.$$

We now show this suffices for the stated KL bound. We follow notation in Lemma 4, whose proof provides a coupling $\Gamma^{(K-\frac{1}{2})}$ between $\mathbf{y}^{(K-1)}$ and $\mathbf{y}_\star^{(K-1)}$ satisfying

$$\mathbb{E}_{(\mathbf{y}^{(K-1)}, \mathbf{y}_\star^{(K-1)})\sim\Gamma^{(K-\frac{1}{2})}}\left[\left\|\mathbf{y}^{(K-1)} - \mathbf{y}_\star^{(K-1)}\right\|_2^2\right] \le 2\eta\epsilon^2.$$

Moreover for every realization of $(\mathbf{y}^{(K-1)}, \mathbf{y}_\star^{(K-1)})$, the derivation in (4) shows

$$D_{\mathrm{KL}}\left(\pi^\star_{\mathbf{y}^{(K-1)}}\|\pi^\star_{\mathbf{y}_\star^{(K-1)}}\right) \le \frac{1}{2\eta^2(\mu + \frac{1}{\eta})}\left\|\mathbf{y}^{(K-1)} - \mathbf{y}_\star^{(K-1)}\right\|_2^2 \le \frac{1}{2\eta}\left\|\mathbf{y}^{(K-1)} - \mathbf{y}_\star^{(K-1)}\right\|_2^2.$$

Finally, applying Lemma 5 with $\xi = (\mathbf{y}^{(K-1)}, \mathbf{y}_\star^{(K-1)})$ and $\pi = \Gamma^{(K-\frac{1}{2})}$ gives the desired claim:

$$D_{\text{KL}}\left(\pi^{(K)} \| \pi^\star\right) \le \mathbb{E}_{(\mathbf{y}^{(K-1)}, \mathbf{y}_\star^{(K-1)}) \sim \Gamma^{(K-\frac{1}{2})}}\left[D_{\text{KL}}\left(\pi^\star_{\mathbf{y}^{(K-1)}} \| \pi^\star_{\mathbf{y}_\star^{(K-1)}}\right)\right] \le \epsilon^2.$$

□

The proof of Theorem 1 given in this section is based on the presentation in [LST21]. Following this result, multiple alternative proofs of Theorem 1 (which extend to more general settings as well) have been found, which we briefly describe here.

In [CCSW22], the step from $\mathbf{x}^{(k)}$ to $\mathbf{y}^{(k)}$ in (2) is viewed as a *forward heat flow*, i.e., the solution to the pure drift equation $\mathrm{d}\mathbf{x}_t = \mathrm{d}\boldsymbol{B}_t$, where $\mathbf{x}^{(k)} = \mathbf{x}_0$ and $\mathbf{y}^{(k)} = \mathbf{x}_\eta$. This results in $\mathbf{y}^{(k)} \sim \mathcal{N}(\mathbf{x}^{(k)}, \eta \mathbf{I}_d)$ as desired by the update (2). On the other hand, the step from $\mathbf{y}^{(k)}$ to $\mathbf{x}^{(k)}$ is viewed as a *backwards heat flow*, i.e., the backwards evolution of Brownian motion conditioned on an endpoint (the state of the stochastic process at time η). We will discuss tools for formalizing this idea in the following lecture on diffusion models, where similar ideas are employed.

Leveraging this perspective allows [CCSW22] to develop fairly explicit formulas for the change in relative entropy after both the forwards and backwards heat flow steps, akin to the entropy decay bound in Lemma 10, Part XVI. Intuitively, the heat flow toolbox provided by [CCSW22] gives a way of deriving a “dynamic” stochastic process that interpolates between the “static” updates (2), which only care about the endpoints of the stochastic process. This gives a way of viewing the RGO as a continuous process, making it more amenable to the stochastic analysis tools from Part XVI and avoiding much of the discretization tedium in Part XVII. Interestingly, via their heat flow proof strategy, [CCSW22] show that Theorem 1 continues to hold under a log-Sobolev constant of $\frac{1}{\mu}$, even when the target is not logconcave. They also give weaker convergence results in χ^2 parameterized by the Poincaré constant, and extensions to other natural functional inequalities.

Finally, [CE22] took this view a step further and developed a general theory for Markov chains induced by *localization processes*. Roughly speaking, in the framework of [CE22], a localization process is any family of (random) densities $\{\pi_t\}_{[0,\eta]}$, such that $\pi_0 = \pi^\star$ is a target stationary density, and $\mathbb{E}\pi_t(\mathbf{x}) = \pi_0(\mathbf{x})$ pointwise on $\mathbf{x} \in \mathbb{R}^d$. This can be viewed as a random density-valued martingale process, that induces a Markov chain by first randomly evolving a current density to time η, and then collapsing it back to time 0 via posterior sampling. Moreover, if all of the densities $\pi_t(\mathbf{x})$ satisfy a functional inequality (e.g., log-Sobolev inequality), then [CE22] use the martingale property to show that the induced Markov chain converges quickly to the target density $\pi^\star$.

By taking the view of $\mathbf{y}^{(k)}$ as the convolution of $\mathbf{x}^{(k)}$ with a noisy channel (i.e., Gaussian), and $\mathbf{x}^{(k+1)}$ as a sample from the updated posterior distribution, [CE22] showed that the RGO is in fact the Markov chain induced by a special localization process called *stochastic localization*, the subject of our final lecture. The improved log-Sobolev constant of the convolved densities due to the Gaussian component (cf. Theorem 4, Part XVI) then yields relative entropy decay.

3 Additional applications

3.1 Structured logconcave sampling

kjtian: TODO: discuss [LST21, FYC23, GLL24].

Log-Lipschitz densities.

Composite densities.

Finite sums.

3.2 General logconcave sampling

kjtian: TODO: discuss [KVZ24, KZ24, KZ25, KV25].

3.3 Non-Euclidean generalizations

kjtian: TODO: discuss [GLL$^+$23, HHBE24].

Source material

Portions of this lecture are based on reference material in [Che24], as well as the author's own experience working in the field.

References

[AC24] Jason M. Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *J. ACM*, 71(3):24, 2024.

[Ana23] Nima Anari. Lecture 19: Continuous sampling. https://nimaanari.com/cs263-autumn2023/assets/files/lecture19.pdf, 2023.

[CCSW22] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory, 2-5 July 2022*, volume 178 of *Proceedings of Machine Learning Research*, pages 2984–3014. PMLR, 2022.

[CE22] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022*, pages 110–122. IEEE, 2022.

[Che24] Sinho Chewi. *Log-Concave Sampling*. 2024.

[FYC23] Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 1473–1521. PMLR, 2023.

[GLL$^+$23] Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Algorithmic aspects of the log-laplace transform and a non-euclidean proximal sampler. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 2399–2439. PMLR, 2023.

[GLL24] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. *J. Priv. Confidentiality*, 14(1), 2024.

[HHBE24] Ye He, Alireza Mousavi Hosseini, Krishnakumar Balasubramanian, and Murat A. Erdogdu. A separation in heavy-tailed sampling: Gaussian vs. stable oracles for proximal samplers. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024.

[KV25] Yunbum Kook and Santosh S. Vempala. Sampling and integration of logconcave functions by algorithmic diffusion. In *STOC '25, 57th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2025.

[KVZ24] Yunbum Kook, Santosh S. Vempala, and Matthew Shunshi Zhang. In-and-out: Algorithmic diffusion for sampling convex bodies. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024.

[KZ24] Yunbum Kook and Matthew S. Zhang. Covariance estimation using markov chain monte carlo. *CoRR*, abs/2410.17147, 2024.

[KZ25] Yunbum Kook and Matthew S. Zhang. Rényi-infinity constrained sampling with d^3 membership queries. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2025*, pages 5278–5306. SIAM, 2025.

[LST21] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 2993–3050. PMLR, 2021.