

# A Conceptual Introduction to Markov Chain Monte Carlo Methods

Joshua S. Speagle

Center for Astrophysics | Harvard & Smithsonian, 60 Garden St., Cambridge,  
MA 02138, USA

[jspeagle@cfa.harvard.edu](mailto:jspeagle@cfa.harvard.edu)

## Аннотация

Методы Монте-Карло с цепью Маркова (МСМС) стали краеугольным камнем многих современных научных анализов, поскольку обеспечивают простой подход к численной оценке неопределенностей в параметрах модели с использованием последовательности случайных выборок. Эта статья представляет собой базовое введение в методы МСМС, устанавливая четкое концептуальное понимание того, какие проблемы пытаются решить методы МСМС, почему мы хотим их использовать и как они работают в теории и на практике. Чтобы развить эти концепции, я излагаю основы байесовского вывода, обсудите, как апостериорные распределения используются на практике, изучите основные подходы к оценке величин, основанных на апостериорных данных, и выведите их связь с выборкой по методу Монте-Карло и МСМС. Затем, используя простую игрушечную задачу, я продемонстрирую, как эти концепции могут быть использованы для понимания преимуществ и недостатков различных подходов МСМС. Упражнения, предназначенные для освещения различных концепций, также включены в статью.

## 1 Introduction

Научный анализ обычно основывается на выводах о лежащих в основе физических моделях из различных источников данных наблюдений. За последние несколько десятилетий качество и количество этих данных существенно возросли, поскольку их сбор и хранение стали быстрее и дешевле. В то же время та же технология, которая позволила собирать огромные объемы данных, также привела к существенному увеличению вычислительной мощности и ресурсов, доступных для их анализа.

В совокупности эти изменения позволили исследовать все более сложные модели с использованием методов, которые может использовать эти вычислительные ресурсы. Это привело к резкому увеличению числа опубликованных работ, основанных на методах Монте-Карло, которые используют комбинацию численного моделирования и генерации случайных чисел для изучения этих моделей.

Одно особенно популярное подмножество методов Монте-Карло известно как Цепочка Маркова Монте-Карло (МСМС). Методы МСМС привлекательны тем, что они предоставляют простой, интуитивно понятный способ как моделировать значения из неизвестного распределения, так и использовать эти смоделированные значения для выполнения последующего анализа. Это позволяет им применяться в самых разных областях.

Благодаря широкому использованию различные обзоры методов МСМС распространены как в рецензируемых, так и в не рецензируемых источниках. В целом, они, как правило, делятся на две группы: статьи, посвященные различным статистическим основам методов МСМС, и статьи, посвященные внедрению и практическому использованию. Читателям, заинтересованным в более подробном ознакомлении с любой из этих тем, рекомендуется ознакомиться с Brooks et al. (2011) и Hogg & Foreman-Mackey (2018) вместе с соответствующими ссылками в них.

Вместо этого в этой статье представлен обзор методов МСМС, направленных на формирование четкого концептуального понимания того, что, почему и как в МСМС, основанного на статистической интуиции. В частности, в нем предпринята попытка систематически ответить на следующие вопросы:

1. Какие проблемы пытаются решить методы МСМС?
2. Почему мы заинтересованы в их использовании?
3. Как они работают в теории и на практике?

Отвечая на эти вопросы, в этой статье обычно предполагается, что читатель в некоторой степени знаком с основами байесовского вывода в теории (например, роль априорных значений) и на практике (например, получение апостериорных значений), базовой статистикой (например, математические ожидания) и основными численными методами (например, суммы Римана). Никаких продвинутых статистических знаний не требуется. Для получения более подробной информации по этим темам, пожалуйста,смотрите Gelman et al. (2013) и Blitzstein & Hwang (2014) вместе с соответствующими ссылками в них.

Схема статьи такова. В §2 я даю краткий обзор байесовского вывода и апостериорных распределений. В §3 я обсуждаю, для чего используются апостериоры на практике,

уделяя особое внимание интеграции и маргинализации. В §4 я описываю базовую схему аппроксимации этих апостериорных интегралов с использованием дискретных сеток. В §5 я иллюстрирую, как методы Монте-Карло становятся естественным продолжением подходов, основанных на сетке. В §6 я обсуждаю, как методы МСМС вписываются в более широкий спектр возможных подходов и их преимущества и недостатки. В §7 я исследую общие проблемы, с которыми сталкиваются методы МСМС. В §8 я исследую, как эти концепции сочетаются на практике, используя простой пример. Я завершаю в §9.

## 2 Bayesian Inference

Во многих научных приложениях у нас есть доступ к некоторым данным  $\mathbf{D}$ , которые мы хотим использовать, чтобы делать выводы об окружающем нас мире. Чаще всего мы хотим интерпретировать эти данные в свете лежащего в их основе *model*  $M$ , который может делать прогнозы относительно данных, которые мы ожидаем увидеть, как функцию некоторых параметров  $\Theta_M$  этой конкретной модели.

Мы можем объединить эти части вместе, чтобы оценить вероятность  $P(\mathbf{D}|\Theta_M, M)$  что мы действительно увидим те данные  $\mathbf{D}$ , которые мы собрали условно (т.е. предполагая) конкретного выбора параметров  $\Theta_M$  из нашей модели  $M$ . Другими словами, если предположить, что наша модель  $M$  верна и параметры  $\Theta_M$  описывают данные, то какова вероятность  $P(\mathbf{D}|\Theta_M, M)$  параметров  $\Theta_M$  на основе наблюдаемых данных  $\mathbf{D}$ ? Предполагая различные значения  $\Theta_M$  дадут различные вероятности, говоря нам о том, какие варианты параметров лучше всего описывают наблюдаемые данные.

В байесовском выводе нас интересует вывод перевернутой величины,  $P(\Theta_M|\mathbf{D}, M)$ . Это описывает вероятность того, что лежащие в основе параметры на самом деле являются  $\Theta_M$ , учитывая наши данные  $\mathbf{D}$  и предполагая определенную модель  $M$ . Используя факторизацию вероятности, мы можем связать эту новую вероятность  $P(\Theta_M|\mathbf{D}, M)$  с вероятностью  $P(\mathbf{D}|\Theta_M, M)$ , описанной выше, как

$$P(\Theta_M|\mathbf{D}, M)P(\mathbf{D}|M) = P(\Theta_M, \mathbf{D}|M) = P(\mathbf{D}|\Theta_M, M)P(\Theta_M|M) \quad (1)$$

где  $P(\Theta_M, \mathbf{D}|M)$  представляет собой совместную вероятность иметь базовый набор параметров  $\Theta_M$ , описывающих данные, и наблюдать конкретный набор данных  $\mathbf{D}$ , которые мы уже собрали.

Перестановка этого равенства в более удобную форму дает нам теорему Бэйса:

$$P(\Theta_M|\mathbf{D}, M) = \frac{P(\mathbf{D}|\Theta_M, M)P(\Theta_M|M)}{P(\mathbf{D}|M)} \quad (2)$$

Теперь это уравнение точно описывает, как наши две вероятности соотносятся друг с другом.

$P(\Theta_M|M)$  часто упоминается как *предшествующий*. Это описывает вероятность наличия определенного набора значений  $\Theta_M$  для нашей данной модели  $M$  до обработки наших данных. Поскольку это не зависит от данных, этот термин часто интерпретируется как представляющий наши ‘предшествующие убеждения’ о том, какими должны

быть  $\Theta_M$  на основе предыдущих измерений, физических проблем и других известных факторов. На практике это приводит к существенному "дополнению" данных другой информацией.

Знаменатель

$$P(\mathbf{D}|M) = \int P(\mathbf{D}|\Theta_M, M)P(\Theta_M|M)d\Theta_M \quad (3)$$

называется доказательством или предельным правдоподобием для нашей модели  $M$  маргинально (т.е. интегрировано) по всем возможным значениям параметров  $\Theta_M$ . В широком смысле, это попытка количественно оценить, насколько хорошо наша модель  $M$  объясняет данные  $\mathbf{D}$  после усреднения по всем возможным значениям  $\Theta_M$  истинных базовых параметров. Другими словами, если наблюдения, предсказанные нашей моделью, похожи на данные  $\mathbf{D}$ , то  $M$  - хорошая модель. Модели, в которых это верно чаще всего, также предпочтительнее моделей, которые дают отличное согласие время от времени, но большую часть времени расходятся. Поскольку в большинстве случаев мы принимаем  $\mathbf{D}$  как данность, это часто оказывается константой.

Наконец,  $P(\Theta_M|\mathbf{D}, M)$  представляет собой наше постериорное распределение. Это количественная оценка нашей веры в  $\Theta_M$  после объединения нашей предварительной интуиции  $P(\Theta_M|M)$  с текущими наблюдениями  $P(\mathbf{D}|\Theta_M, M)$  и нормализации по общему доказательству  $P(\mathbf{D}|M)$ . Постериор будет представлять собой некоторый компромисс между предшествующим и вероятностью, причем точное сочетание зависит от силы и свойств предшествующего и качества данных, используемых для получения вероятности. Схематичная иллюстрация показана на [Figure 1](#).

Во всей остальной части статьи я буду писать эти четыре термина (вероятность, предшествующее (приор), доказательство, последующее (апостериор)), используя сокращенные обозначения, такие как

$$\mathcal{P}(\Theta) \equiv \frac{\mathcal{L}(\Theta)\pi(\Theta)}{\int \mathcal{L}(\Theta)\pi(\Theta)d\Theta} \equiv \frac{\mathcal{L}(\Theta)\pi(\Theta)}{\mathcal{Z}} \quad (4)$$

где  $\mathcal{P}(\Theta) \equiv P(\Theta_M|\mathbf{D}, M)$  - апостериор,  $\mathcal{L}(\Theta) \equiv P(\mathbf{D}|\Theta_M, M)$  - вероятность,  $\pi(\Theta) \equiv P(\Theta_M|M)$  - приоритет, и константа  $\mathcal{Z} \equiv P(\mathbf{D}|M)$  - доказательство. Для удобства я опустил обозначения модели  $M$  и данных  $\mathbf{D}$ , поскольку в большинстве случаев данные и модель считаются фиксированными, но при необходимости я буду вводить их снова.

Прежде чем продолжить, я хотел бы подчеркнуть, что интерпретация любого результата хороша лишь настолько, насколько хороши модели и приорные оценки, которые лежат в их основе. Попытки исследовать последствия той или иной модели с помощью, например, некоторых методов, описанных в этой статье, по сути, являются второстепенной задачей по сравнению с построением разумной модели с хорошо мотивированными приматами в первую очередь. Я настоятельно рекомендую читателям помнить об этой идее на протяжении всей оставшейся части этой работы.

Упражнение: Среднее значение шума

Настройка

Рассмотрим случай, когда у нас есть станции мониторинга температуры, расположенные по всему городу. Каждая станция  $i$  проводит зашумленное измерение  $\hat{T}_i$  темпера-

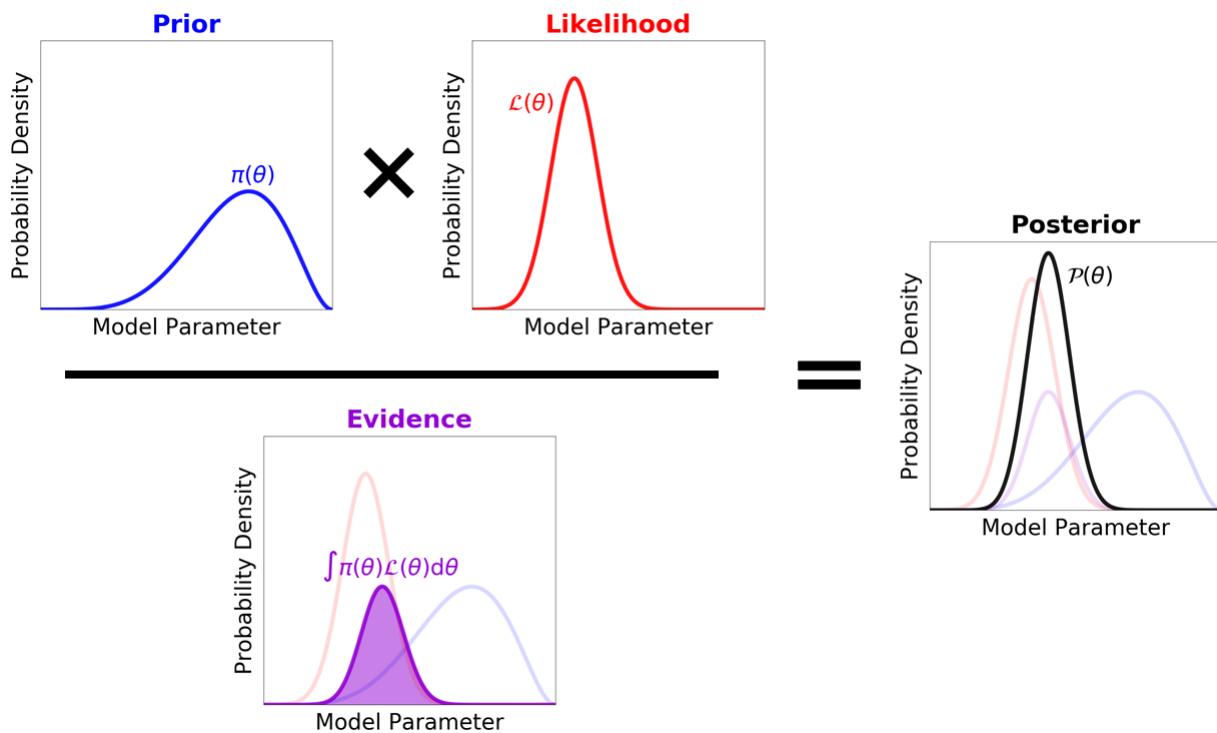


Рис. 1: Иллюстрация теоремы Байеса. Постериорная вероятность  $\mathcal{P}(\Theta)$  (черный) параметров нашей модели  $\Theta$  основана на комбинации наших предварительных убеждений  $\pi(\Theta)$  (синий) и вероятности  $\mathcal{L}(\Theta)$  (красный), нормированной на общее доказательство  $\mathcal{Z} = \int \pi(\Theta)\mathcal{L}(\Theta)d\Theta$  (фиолетовый) для нашей конкретной модели. Дополнительные подробности см. в §2.

туры в любой день с некоторым шумом измерения  $\sigma_i$ . Мы будем считать, что наши измерения  $\hat{T}_i$  следуют нормальному (т.е. гауссовскому) распределению со средним  $T$  и стандартным отклонением  $\sigma_i$ , таким, что

$$\hat{T}_i \sim \mathcal{N}[T, \sigma_i]$$

Это означает, что вероятность

$$P(\hat{T}_i|T, \sigma_i) \equiv \mathcal{N}[T, \sigma_i] = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2} \frac{(\hat{T}_i - T)^2}{\sigma_i^2}\right]$$

для каждого наблюдения и

$$P(\{\hat{T}_i\}_{i=1}^n|T, \{\sigma_i\}_{i=1}^n) = \prod_{i=1}^n P(\hat{T}_i|T, \sigma_i)$$

для набора из  $n$  наблюдений.

Предположим, что у нас есть пять независимых шумных измерений температуры (в градусах Цельсия) с нескольких станций мониторинга

$$\hat{T}_1 = 26.3, \hat{T}_2 = 30.2, \hat{T}_3 = 29.4, \hat{T}_4 = 30.1, \hat{T}_5 = 29.8$$

с соответствующими неопределенностями

$$\sigma_1 = 1.7, \sigma_2 = 1.8, \sigma_3 = 1.2, \sigma_4 = 0.5, \sigma_5 = 1.3$$

Рассматривая исторические данные, мы обнаруживаем, что типичная базовая температура  $T$  в течение подобных дней имеет примерно нормальное распределение со средним значением  $T_{\text{prior}} = 25$  и вариацией  $\sigma_{\text{prior}} = 1.5$ :

$$T \sim \mathcal{N}[T_{\text{prior}} = 25, \sigma_{\text{prior}} = 1.5]$$

Постановка проблемы

Используя эти предположения, вычислите:

1. приор  $\pi(T)$ ,
2. вероятность  $\mathcal{L}(T)$ , и
3. апостериор  $\mathcal{P}(T)$ .

учитывая наши наблюдаемые данные  $\{\hat{T}_i\}$  и ошибки  $\{\sigma_i\}$  в диапазоне температур  $T$ . Как различаются эти три условия? Похоже ли предварительное предположение на хорошее? Почему или почему нет?

### 3 Для чего нужны апостериорные распределения?

Выше я описал, как теорема Байеса способна объединить наши предварительные убеждения и наблюдаемые данные в новую апостериорную оценку  $\mathcal{P}(\Theta) \propto \mathcal{L}(\Theta)\pi(\Theta)$ . Однако это только половина проблемы. Получив апостериор, мы должны затем использовать его, чтобы сделать выводы об окружающем нас мире. В целом, способы, которыми мы хотим использовать апостериоры, делятся на несколько широких категорий:

1. Сделать обоснованные предположения: сделать обоснованное предположение о том, какие параметры лежат в основе модели.
2. Квантование неопределенности: определить ограничения на диапазон возможных значений параметров модели.
3. Генерирование прогнозов: предельная оценка неопределенности параметров модели для предсказания наблюдаемых или других переменных, зависящих от параметров модели.
4. Сравнение моделей: использование доказательств, полученных с помощью различных моделей, для определения того, какая модель более благоприятна.

Для достижения этих целей нам часто интереснее попытаться использовать апостериор для оценки различных ограничений на сами параметры  $\Theta$  или другие величины  $f(\Theta)$ , которые могут быть основаны на них. Это часто зависит от маргинализации неопределенностей, характеризуемых нашим апостериором (через правдоподобие и приоритет). Доказательство  $\mathcal{Z}$ , например, снова является просто интегралом правдоподобия и предшествования по всем возможным параметрам:

$$\mathcal{Z} = \int \mathcal{L}(\Theta)\pi(\Theta)d\Theta \equiv \int \tilde{\mathcal{P}}(\Theta)d\Theta \quad (5)$$

где  $\tilde{\mathcal{P}}(\Theta) \equiv \mathcal{L}(\Theta)\pi(\Theta)$  это ненормированный апостериор.

Аналогично, если мы изучаем поведение подмножества "интересных" параметров  $\Theta_{\text{int}}$  из  $\Theta = \{\Theta_{\text{int}}, \Theta_{\text{nuis}}\}$ , мы хотим маргинализировать поведение оставшихся "неприятных" параметров  $\Theta_{\text{nuis}}$ , чтобы увидеть, как они могут повлиять на  $\Theta_{\text{int}}$ . Этот процесс довольно прост, если известно все апостериорное значение  $\Theta$ :

$$\mathcal{P}(\Theta_{\text{int}}) = \int \mathcal{P}(\Theta_{\text{int}}, \Theta_{\text{nuis}}) d\Theta_{\text{nuis}} = \int \mathcal{P}(\Theta) d\Theta_{\text{nuis}} \quad (6)$$

Другие величины, как правило, могут быть получены из значения ожидания различных функций  $f(\Theta)$ , зависящих от параметров, по отношению к апостериору:

$$\mathbb{E}_{\mathcal{P}} [f(\Theta)] \equiv \frac{\int f(\Theta)\mathcal{P}(\Theta)d\Theta}{\int \mathcal{P}(\Theta)d\Theta} = \frac{\int f(\Theta)\tilde{\mathcal{P}}(\Theta)d\Theta}{\int \tilde{\mathcal{P}}(\Theta)d\Theta} = \int f(\Theta)\mathcal{P}(\Theta)d\Theta \quad (7)$$

поскольку  $\int \mathcal{P}(\Theta)d\Theta = 1$  по определению и  $\tilde{\mathcal{P}}(\Theta) \propto \mathcal{P}(\Theta)$ . Это представляет собой средневзвешенное значение  $f(\Theta)$ , где при каждом значении  $\Theta$  мы взвешиваем полученное значение  $f(\Theta)$ , исходя из вероятности того, что это значение является правильным.

В совокупности мы видим, что почти во всех случаях нам интереснее вычислять интегралы по апостериору, чем знать сам апостериор. Другими словами, апостериор редко бывает полезен сам по себе; в основном он становится полезным при интегрировании по нему.

Это различие между оценкой ожиданий и других интегралов по апостериору и оценкой апостериора как такового является ключевым элементом байесовского вывода. Это различие имеет огромное значение, когда дело доходит до практического выполнения выводов, поскольку часто бывает так, что мы можем получить отличную оценку  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$ , даже если у нас крайне плохая оценка  $\mathcal{P}(\Theta)$  или  $\tilde{\mathcal{P}}(\Theta)$ .

Ниже приводится более подробная информация, чтобы проиллюстрировать, как конкретные категории, описанные выше, превращаются в конкретные интегралы по (ненормированному) заднему числу. Пример показан на [Figure 2](#).

### 3.1 Делаем обоснованные предположения

Один из основных постулатов байесовского вывода состоит в том, что мы не знаем ни истинной модели  $M_*$ , ни ее истинных базовых параметров  $\Theta_*$ , характеризующих наблюдаемые данные: имеющаяся у нас модель  $M$  почти всегда является упрощением того, что происходит на самом деле. Однако если мы предположим, что наша текущая модель  $M$  верна, то мы можем попытаться использовать наш апостериор  $\mathcal{P}(\Theta)$ , чтобы предложить точечную оценку  $\hat{\Theta}$ , которая, по нашему мнению, является довольно хорошим предположением для истинного значения  $\Theta_*$ .

Что именно считается “хорошим”? Это зависит от того, что именно нас волнует. В общем случае, мы можем оценить “хорошо”, задав противоположный вопрос: насколько сильно мы наказаны, если наша оценка  $\hat{\Theta} \neq \Theta_*$  окажется неверной? Часто для этого используется функция потерь  $L(\hat{\Theta}|\Theta_*)$ , которая наказывает нас, когда наша точечная оценка  $\hat{\Theta}$  отличается от  $\Theta_*$ . Примером общей функции потерь является  $L(\hat{\Theta}|\Theta_*) = |\hat{\Theta} - \Theta_*|^2$  (т.е. квадратичная потеря), где неправильное предположение наказывается квадратом величины расхождения между предположением  $\hat{\Theta}$  и истинным значением  $\Theta_*$ .

К сожалению, мы не знаем, каково реальное значение  $\Theta_*$ , чтобы оценить истинный проигрыш. Однако мы можем поступить следующим образом и вычислить ожидаемый убыток, усредненный по всем возможным значениям  $\Theta_*$ , основываясь на нашем апостериоре:

$$L_{\mathcal{P}}(\hat{\Theta}) \equiv \mathbb{E}_{\mathcal{P}} [L(\hat{\Theta}|\Theta)] = \int L(\hat{\Theta}|\Theta) \mathcal{P}(\Theta) d\Theta \quad (8)$$

Тогда разумным выбором для  $\hat{\Theta}$  будет значение, которое минимизирует этот ожидаемый убыток вместо фактического (неизвестного) убытка:

$$\hat{\Theta} \equiv \underset{\Theta'}{\operatorname{argmin}} [L_{\mathcal{P}}(\Theta')] \quad (9)$$

где  $\operatorname{argmin}$  указывает на значение (аргумент)  $\Theta'$ , которое минимизирует ожидаемый убыток  $L_{\mathcal{P}}(\Theta')$ .

Хотя эта стратегия может работать для любой произвольной функции потерь, решение  $\hat{\Theta}$  часто требует использования численных методов и повторного интегрирования по

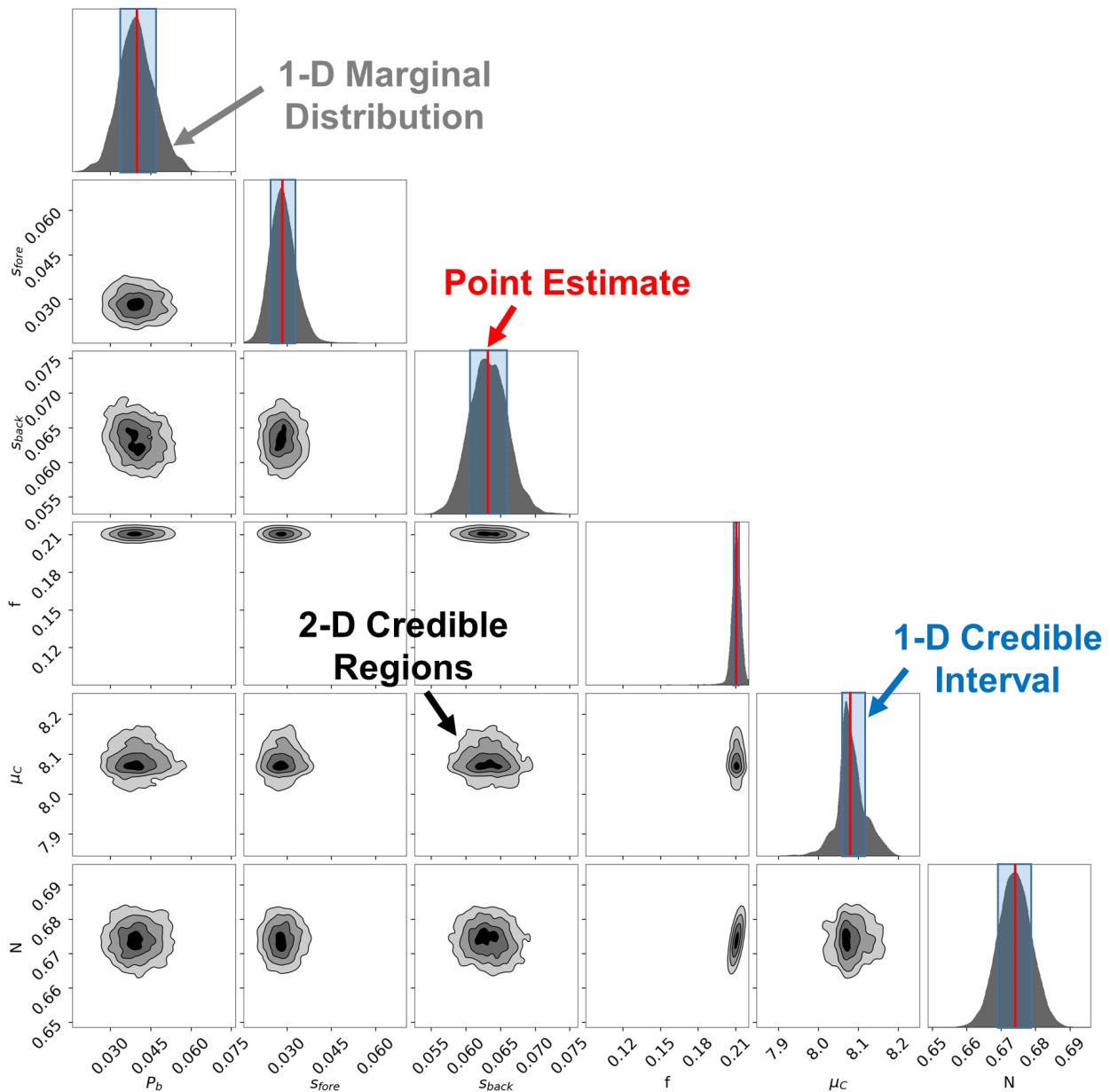


Рис. 2: Угловой график, показывающий пример практического использования астериоров. На каждой из верхних панелей показано одномерное маргинальное апостериорное распределение для каждого параметра (серый), а также соответствующие медианные точечные оценки (красный) и 68%-ные доверительные интервалы (синий). На каждой центральной панели показаны 10%, 40%, 65% и 85% доверительные области для каждого двумерного маргинального апостериорного распределения. Дополнительные сведения см. в разделе §3.

$\mathcal{P}(\Theta)$ . Однако для определенных функций потерь существуют аналитические решения. Например, легко показать (и это будет интересным упражнением для заинтересованного читателя), что оптимальная оценка точки  $\hat{\Theta}$  при квадратичных потерях - это просто среднее.

### 3.2 Quantifying Uncertainty

Во многих случаях нас интересует не просто вычисление предсказания  $\hat{\Theta}$  для  $\Theta_*$ , но и ограничение области  $\mathcal{C}(\Theta)$  возможных значений, внутри которой  $\Theta_*$  может лежать с некоторой долей уверенности. Другими словами, можем ли мы построить область  $\mathcal{C}_X$  такую, что мы считаем, что существует  $X\%$  шанс, что она содержит  $\Theta_*$ ?

Существует множество возможных определений этой вероятной области. Одно из общих определений - это область выше некоторого порога апостериорности  $\mathcal{P}_X$ , в которой содержится  $X\%$  апостериорности, т.е. где

$$\int_{\Theta \in \mathcal{C}_X} \mathcal{P}(\Theta) d\Theta = \frac{X}{100} \quad (10)$$

с учетом

$$\mathcal{C}_X \equiv \{\Theta : \mathcal{P}(\Theta) \geq \mathcal{P}_X\} \quad (11)$$

Другими словами, мы хотим проинтегрировать наш астериор по всем  $\Theta$ , где значение  $\mathcal{P}(\Theta) > \mathcal{P}_X$  больше некоторого порога  $\mathcal{P}_X$ , где  $\mathcal{P}_X$  задается так, чтобы этот интеграл охватывал  $X\%$  от полного астериора. Обычно для  $X$  выбирают 68% и 95% (т.е. "1-сигма" и "2-сигма" доверительные интервалы).

В частном случае, когда наше (маргинальное) апостериори является одномерным, достоверные интервалы часто определяются с помощью перцентилей, а не порогов, где место  $x_p$  расположения  $p$ -го перцентиля определяется как

$$\int_{-\infty}^{x_p} \mathcal{P}(x) dx = \frac{p}{100} \quad (12)$$

Мы можем использовать их для определения достоверной области  $[x_{\text{low}}, x_{\text{high}}]$ , содержащей  $Y\%$  данных, взяв  $x_{\text{low}} = x_{(1-Y)/2}$  и  $x_{\text{high}} = x_{(1+Y)/2}$ . Хотя это приводит к асимметричным пороговым значениям и не обобщается на более высокие размерности, преимуществом этого метода является то, что он всегда охватывает медианное значение  $x_{50}$  и имеет равные хвостовые вероятности (т. е.  $(1 - Y)/2\%$  апостериорного значения с каждой стороны).

В целом, когда в тексте упоминаются "достоверные интервалы", следует исходить из определения перцентиля, если явно не указано иное.

### 3.3 Предсказание

Помимо попыток оценить основные параметры нашей модели, мы также часто хотим сделать предсказания других наблюдений или переменных, которые зависят от параметров нашей модели. Если мы считаем, что знаем истинные базовые параметры модели

$\Theta_*$ , то этот процесс прост. Однако, учитывая, что у нас есть доступ только к апостериорному распределению  $P(\Theta)$  по возможным значениям  $\Theta_*$ , чтобы предсказать, что произойдет, нам нужно сделать маргинализацию на эту неопределенность.

Мы можем количественно выразить эту интуицию с помощью апостериорного прогноза  $P(\tilde{\mathbf{D}}|\mathbf{D})$ , который представляет собой вероятность увидеть новые данные  $\tilde{\mathbf{D}}$  на основе имеющихся данных  $\mathbf{D}$ :

$$P(\tilde{\mathbf{D}}|\mathbf{D}) \equiv \int P(\tilde{\mathbf{D}}|\Theta)P(\Theta|\mathbf{D})d\Theta \equiv \int \tilde{\mathcal{L}}(\Theta)\mathcal{P}(\Theta)d\Theta = \mathbb{E}_{\mathcal{P}} [\tilde{\mathcal{L}}(\Theta)] \quad (13)$$

Другими словами, для гипотетических данных  $\tilde{\mathbf{D}}$  мы хотим вычислить ожидаемое значение вероятности  $\tilde{\mathcal{L}}(\Theta)$  по всем возможным значениям  $\Theta$  на основе текущего апостериорного  $\mathcal{P}(\Theta)$ .

### 3.4 Сравнение моделей

Последний момент, представляющий интерес во многих байесовских анализах, - это попытка выяснить, благоприятствуют ли данные какой-либо модели (моделям), которую мы предполагаем в нашем анализе. Наш выбор приора или конкретный способ параметризации данных может привести к существенным различиям в интерпретации результатов.

Мы можем сравнить две модели, вычислив коэффициент Bayes factor:

$$\mathcal{R}_2^1 \equiv \frac{P(M_1|\mathbf{D})}{P(M_2|\mathbf{D})} = \frac{P(\mathbf{D}|M_1)P(M_1)}{P(\mathbf{D}|M_2)P(M_2)} \equiv \frac{\mathcal{Z}_1 \pi_1}{\mathcal{Z}_2 \pi_2} \quad (14)$$

где  $\mathcal{Z}_M$  - снова доказательства в пользу модели  $M$ , а  $\pi_M$  - наша предварительная вера в то, что  $M$  верна по сравнению с конкурирующей моделью. В совокупности, фактор Байеса  $\mathcal{R}$  говорит нам, насколько конкретная модель предпочтительнее другой, учитывая наблюдаемые данные, предельные значения всех возможных параметров модели  $\Theta_M$  и нашу предыдущую относительную уверенность в модели.

Еще раз отметим, что вычисление  $\mathcal{Z}_M$  требует вычисления интеграла  $\int \tilde{\mathcal{P}}(\Theta)d\Theta$  от ненормированного заднего числа  $\tilde{\mathcal{P}}(\Theta)$  по  $\Theta$ . В сочетании с другими примерами, описанными в этом разделе, становится ясно, что многие распространенные случаи использования байесовского анализа основаны на вычислении интегралов по (возможно, ненормированному) апостериору.

### Упражнение: Пересмотр шумного значения

#### Setup

Вернемся к нашему температурному апостериору  $\mathcal{P}(T)$  из §2. Мы хотим использовать этот результат для получения интересных оценок и ограничений на возможную базовую температуру  $T$ .

### Точечные оценки

mean можно определить как точечную оценку  $\hat{\Theta}$ , которая минимизирует ожидаемые потери  $L_{\mathcal{P}}(\hat{\Theta})$  при квадратичных потерях:

$$L_{\text{mean}}(\hat{\Theta}|\Theta_*) = |\hat{\Theta} - \Theta_*|^2$$

median может быть определена как точечная оценка, которая минимизирует  $L_{\mathcal{P}}(\hat{\Theta})$  при абсолютных потерях:

$$L_{\text{med}}(\hat{\Theta}|\Theta_*) = |\hat{\Theta} - \Theta_*|$$

mode можно определить как точечную оценку, которая минимизирует  $L_{\mathcal{P}}(\hat{\Theta})$  при ‘катастрофических’ потерях:

$$L_{\text{mode}}(\hat{\Theta}|\Theta_*) = -\delta(|\hat{\Theta} - \Theta_*|)$$

где  $\delta(\cdot)$  - дельта-функция Дирака, определенная так, что

$$\int f(x)\delta(x-a)dx = f(a)$$

Учитывая эти выражения для среднего, медианы и моды, оцените соответствующие оценки температурных точек  $T_{\text{mean}}$ ,  $T_{\text{med}}$  и  $T_{\text{mode}}$  из нашего соответствующего постера. Не стесняйтесь экспериментировать с различными аналитическими и численными методами для выполнения этих расчетов.

Мы можем ожидать, что исторические данные, которые мы использовали для наших приор, могут не так хорошо работать сегодня, если произошли некоторые долгосрочные изменения в средней температуре. Например, мы ожидаем, что средняя температура со временем увеличилась, и поэтому мы, возможно, не захотим штрафовать более жаркие температуры  $T \geq T_{\text{prior}}$  так же сильно, как более холодные  $T < T_{\text{prior}}$ . Мы можем закодировать эту информацию в асимметричной функции потерь, например

$$L(\hat{T}|T_*) = \begin{cases} |\hat{T} - T_*|^3 & T < T_{\text{prior}} \\ |\hat{T} - T_*| & T \geq T_{\text{prior}} \end{cases}$$

Какова оптимальная точечная оценка  $T_{\text{asym}}$ , которая минимизирует ожидаемые потери в этом случае?

### Достоверные интервалы

Далее попробуем количественно оценить неопределенность. Учитывая апостериор  $\mathcal{P}(T)$ , вычислите 50%, 80% и 95% доверительных интервалов, используя апостериорные пороги  $\mathcal{P}_X$ . Затем вычислите эти доверительные интервалы с помощью перцентилей. Есть ли различия между доверительными интервалами, вычисленными двумя методами? Почему или почему нет?

## Апостериорное предсказание

Чтобы распространить наши неопределенности на следующие наблюдения, вычислим апостериорное предсказание  $P(\hat{T}_6 | \{\hat{T}_1, \dots, \hat{T}_5\})$  по диапазону возможных измерений температуры  $\hat{T}_6$  для следующих наблюдений с учетом предыдущих пяти  $\{\hat{T}_1, \dots, \hat{T}_5\}$ , предполагая неопределенность  $\sigma_6 = 0$ ,  $\sigma_6 = 0.5$ , и  $\sigma_6 = 2$ .

## Сравнение моделей

Наконец, мы хотим выяснить, является ли наша предварительная оценка хорошим предположением. Используя численные методы, вычислите доказательство  $\mathcal{Z}$  для нашего приоритета по умолчанию со средним  $T_{\text{prior}} = 25$  и стандартным отклонением  $\sigma_{\text{prior}} = 1.5$ . Затем сравните их с доказательствами, полученными на основе альтернативного приоритета, где мы предполагаем, что температура выросла примерно на пять Сравнение моделей градусов со средним  $T_{\text{prior}} = 30$ , но с соответствующей большей неопределенностью  $\sigma_{\text{prior}} = 3$ . Является ли одна модель особенно предпочтительной по сравнению с другой?

## 4 Аппроксимация апостериорных интегралов с помощью сеток

Теперь я хочу изучить методы оценки апостериорных интегралов. Хотя в некоторых случаях (например, в случае сопряженных приоров) их можно вычислить аналитически, в общем случае это не так. Поэтому для правильной оценки величин, подобных тем, что описаны в §3, необходимо использовать численные методы (освещенные в предыдущих упражнениях).

Для начала я рассмотрю случай, когда наш интеграл по  $\Theta$  является одномерным. В этом случае мы можем аппроксимировать его с помощью стандартных численных методов, таких как сумма Римана по дискретной сетке точек:

$$\mathbb{E}_{\mathcal{P}} [f(\Theta)] = \int f(\Theta) \mathcal{P}(\Theta) d\Theta \approx \sum_{i=1}^n f(\Theta_i) \mathcal{P}(\Theta_i) \Delta\Theta_i \quad (15)$$

где

$$\Delta\Theta_i = \Theta_{j+1} - \Theta_j \quad (16)$$

это просто расстояние между множеством точек  $j = 1, \dots, n + 1$  на базовой сетке и

$$\Theta_i = \frac{\Theta_{j+1} + \Theta_j}{2} \quad (17)$$

определяется как средняя точка между это просто расстояние между множеством точек  $j = 1, \dots, n + 1$  на базовой сетке иен  $\Theta_j$  и  $\Theta_{j+1}$ .<sup>1</sup> Как показано на Figure 3, этот подход сродни попытке аппроксимировать интеграл с помощью дискретного набора  $n$  прямоугольников с высотой  $f(\Theta_i) \mathcal{P}(\Theta_i)$  и шириной  $\Delta\Theta_i$ .

<sup>1</sup>Выбор  $\Theta_i$  в качестве одной из конечных точек дает последовательное поведение (см. §4.3) при увеличении числа точек сетки  $n \rightarrow \infty$ , но обычно приводит к большим погрешностям при конечном  $n$ .

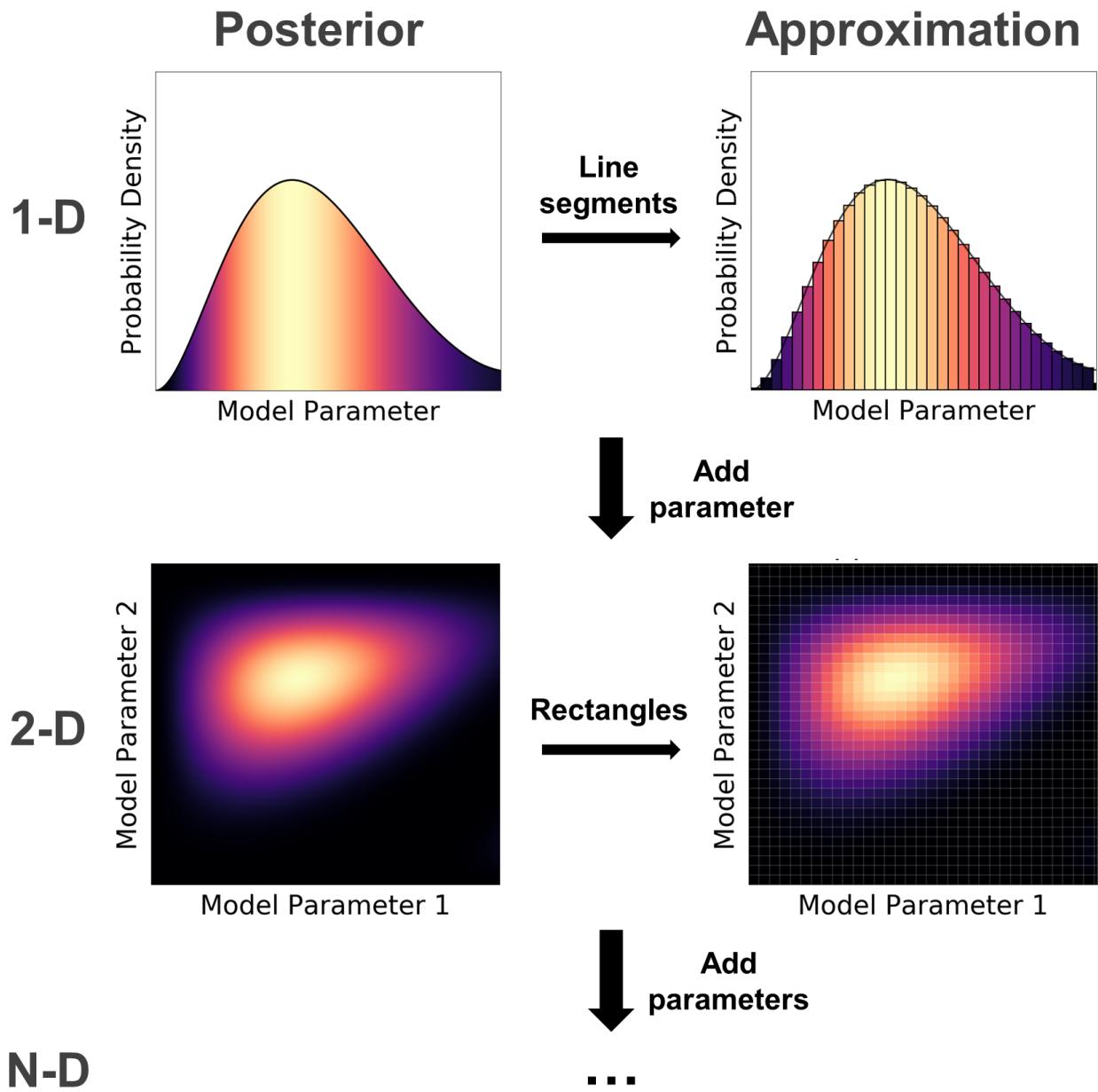


Рис. 3: Иллюстрация того, как аппроксимировать апостериорные интегралы с помощью дискретной сетки точек. Мы разбиваем апостериор на смежные области, определяемые позицией  $\Theta_i$  (например, конечной или средней точкой) с соответствующей плотностью  $\mathcal{P}(\Theta_i)$  и объемом  $\Delta\Theta_i$  по сетке с  $i = 1, \dots, n$  элементами. Наш интеграл может быть аппроксимирован сложением каждой из этих областей пропорционально задней массе  $\mathcal{P}(\Theta_i) \times \Delta\Theta_i$ , содержащейся в ней. В одномерном пространстве (вверху) эти элементы объема  $\Delta\Theta_i$  соответствуют отрезкам прямых, а в двухмерном (в середине) - прямоугольникам. Это можно обобщить на более высокие измерения (внизу), где мы вместо этого использовали N-D кубоиды. Дополнительные подробности см. в разделе §4.

Эту идею можно обобщить на более высокие измерения. В этом случае вместо разбиения интеграла на  $n$  одномерных сегментов мы можем разложить его на множество  $n$  N-D кубоидов. Тогда вклад каждого из этих кубоидов пропорционален произведению "высоты"  $f(\Theta_i)\mathcal{P}(\Theta_i)$  и объема

$$\Delta\Theta_i = \prod_{j=1}^d \Delta\Theta_{i,j} \quad (18)$$

где  $\Delta\Theta_{i,j}$  - ширина  $i$ -го кубоида в  $j$ -м измерении. См. [Figure 3](#) для наглядного представления этой процедуры.

Подставив  $\mathcal{P}(\Theta) = \tilde{\mathcal{P}}(\Theta)/\mathcal{Z}$  в значение ожидания и заменив все интегралы их приближениями на основе сетки, получим:

$$\mathbb{E}_{\mathcal{P}}[f(\Theta)] = \frac{\int f(\Theta)\mathcal{P}(\Theta)d\Theta}{\int \mathcal{P}(\Theta)d\Theta} = \frac{\int f(\Theta)\tilde{\mathcal{P}}(\Theta)d\Theta}{\int \tilde{\mathcal{P}}(\Theta)d\Theta} \approx \frac{\sum_{i=1}^n f(\Theta_i)\tilde{\mathcal{P}}(\Theta_i)\Delta\Theta_i}{\sum_{i=1}^n \tilde{\mathcal{P}}(\Theta_i)\Delta\Theta_i} \quad (19)$$

Обратите внимание, что знаменатель теперь представляет собой оценку доказательств:

$$\mathcal{Z} = \int \tilde{\mathcal{P}}(\Theta)d\Theta \approx \sum_{i=1}^n \tilde{\mathcal{P}}(\Theta_i)\Delta\Theta_j \quad (20)$$

Эта замена ненормированного апостериорного  $\tilde{\mathcal{P}}(\Theta)$  на апостериорное  $\mathcal{P}(\Theta)$  является важной частью вычисления ожидаемых значений на практике, поскольку мы можем вычислить  $\tilde{\mathcal{P}}(\Theta) = \mathcal{L}(\Theta)\pi(\Theta)$  непосредственно без знания  $\mathcal{Z}$ .

## 4.1 Проклятие размерности

Хотя этот подход прост, он имеет один непосредственный и серьезный недостаток: общее количество точек сетки увеличивается экспоненциально с ростом числа измерений. Например, если предположить, что у нас есть примерно  $k \geq 2$  точек сетки в каждом измерении, то общее количество точек  $n$  в нашей сетке увеличивается как

$$n \sim \prod_{j=1}^d k = k^d \quad (21)$$

Это означает, что даже в абсолютном лучшем случае, когда  $k = 2$ , мы имеем масштабирование  $2^d$ .

Это ужасное масштабирование часто называют проклятием размерности. Эта экспоненциальная зависимость оказывается общим свойством высокоразмерных распределений (т.е. апостериоров моделей с большим числом параметров), к которому я вернусь позже в §7.

## 4.2 Эффективный размер выборки

Помимо экспоненциального масштабирования размерности, у использования сеток есть и более тонкий недостаток. Поскольку мы не знаем форму распределения заранее, вклад

каждой части сетки (т.е. каждого N-D кубоида) может быть крайне неравномерным в зависимости от структуры сетки. Другими словами, эффективность этого подхода зависит не только от количества точек сетки  $n$ , но и от места их распределения. Если мы плохо определим точки сетки, мы можем получить много точек, расположенных в областях, где  $\tilde{\mathcal{P}}(\Theta)$  и/или  $f(\Theta)\tilde{\mathcal{P}}(\Theta)$  относительно малы. Это означает, что в их соответствующих суммах будет доминировать небольшое количество точек с гораздо большими относительными "весами". В идеале мы должны увеличить разрешение сетки в тех областях, где апостериорное значение велико, и уменьшить его в других местах, чтобы смягчить этот эффект.

Обратите внимание, что мы используем термин "веса" в предыдущем абзаце вполне осознанно. Если вспомнить нашу первоначальную аппроксимацию, то форма уравнения (19) очень похожа на ту, которая может быть использована для вычисления взвешенного выборочного среднего для  $f(\Theta)$ . В этом случае, когда у нас есть  $n$  наблюдений  $\{f_1, \dots, f_n\}$  с соответствующими весами  $\{w_1, \dots, w_n\}$ , взвешенное среднее находится просто:

$$\hat{f}_{\text{mean}} \equiv \frac{\sum_{i=1}^n w_i f_i}{\sum_{i=1}^n w_i} \quad (22)$$

Действительно, если мы определим

$$f_i \equiv f(\Theta_i), \quad w_i \equiv \tilde{\mathcal{P}}(\Theta_i)\Delta\Theta_i \quad (23)$$

тогда связь между взвешенным выборочным средним в уравнении (22) и матожиданием от нашей сетки в уравнении (19) становится явной:

$$\mathbb{E}_{\mathcal{P}} [f(\Theta)] \approx \frac{\sum_{i=1}^n f(\Theta_i)\tilde{\mathcal{P}}(\Theta_i)\Delta\Theta_i}{\sum_{i=1}^n \tilde{\mathcal{P}}(\Theta_i)\Delta\Theta_i} \equiv \frac{\sum_{i=1}^n w_i f_i}{\sum_{i=1}^n w_i} \quad (24)$$

Рассматривая нашу сетку как набор образцов  $n$ , мы также можем рассмотреть соответствующий эффективный размер выборки (ESS)  $n_{\text{eff}} \leq n$ . В ESS заложена идея о том, что не все наши выборки несут одинаковое количество информации: если у нас есть  $n$  образцов, которые очень похожи друг на друга, мы ожидаем получить значительно худшую оценку, чем если у нас есть  $n$  образцов, которые сильно отличаются друг от друга. Это происходит потому, что информация в коррелированных выборках, по крайней мере, частично избыточна по отношению друг к другу, причем количество избыточности увеличивается с ростом силы корреляции: в то время как две независимые выборки предоставляют совершенно уникальную информацию о распределении и никакой информации друг о друге, две коррелированные выборки вместо этого представляют некоторую информацию друг о друге за счет основного распределения.

Возвращаясь к сеткам, это соответствие означает, что теоретически мы можем получить оценку матожидания  $\mathbb{E}_{\mathcal{P}} [f(\Theta)]$ , которая будет по крайней мере столь же хороша, как и та, которую мы могли бы иметь в настоящее время, используя меньшее число  $n_{\text{eff}} \leq n$  точек сетки, если бы мы могли распределить их более эффективно. Это различие имеет значение, потому что ошибки в нашей оценке матожидания обычно масштабируются как функция  $n_{\text{eff}}$ , а не  $n$ . Например, ошибка среднего значения обычно составляет  $\propto n_{\text{eff}}^{-1/2}$ , а не  $\propto n^{-1/2}$ .



Рис. 4: Пример того, как изменение расстояния (элементов объема) сетки может кардинально повлиять на связанную с ней оценку апостериорных интегралов. На игрушечном двухмерном апостериорном интеграле  $\mathcal{P}(\Theta)$  простое изменение расстояния между элементами соответствующей двухмерной сетки  $30 \times 30$  резко влияет на эффективный размер выборки (ESS) (см. §4.2). Различия между плохим расстоянием (слева), равномерным расстоянием (в середине) и оптимальным расстоянием (справа) приводят к разнице в ESS на порядок величины, что видно по распределению весов (внизу), связанных с элементами объема каждой сетки. Дополнительные подробности см. в §4.2.

Мы можем количественно оценить идеи, лежащие в основе ESS, как обсуждалось выше, введя формальное определение, следуя Kish (1965):

$$n_{\text{eff}} \equiv \frac{\left(\sum_{i=1}^n w_i\right)^2}{\sum_{i=1}^n w_i^2} \quad (25)$$

В соответствии с нашей интуицией, наилучшим случаем при таком определении является тот, в котором все веса равны ( $w_i = w$ ):

$$n_{\text{eff}}^{\text{best}} = \frac{\left(\sum_{i=1}^n w_i\right)^2}{\sum_{i=1}^n w_i^2} = \frac{(nw)^2}{\sum_{i=1}^n w^2} = \frac{n^2 w^2}{nw^2} = n \quad (26)$$

Аналогичным образом, наихудшим случаем является тот, когда весь вес сосредоточен вокруг одной выборки ( $w_i = w$  для  $i = j$  и  $w_i = 0$  в противном случае):

$$n_{\text{eff}}^{\text{worst}} = \frac{\left(\sum_{i=1}^n w_i\right)^2}{\sum_{i=1}^n w_i^2} = \frac{(w)^2}{w^2} = 1 \quad (27)$$

В первом случае (при  $n_{\text{eff}}^{\text{best}}$ ) каждый из элементов нашей сетки вносит примерно одинаковый вклад в интеграл, а во втором (при  $n_{\text{eff}}^{\text{worst}}$ ) весь интеграл по существу содержится только в одной из областей N-D кубоида  $n$ . Иллюстрация такого поведения показана на Figure 4.

### 4.3 Сходимость и согласованность

Теперь, когда я обрисовал взаимосвязь между структурой нашей сетки и ESS, я хочу рассмотреть два последних вопроса: сходимость и согласованность. Сходимость - это идея о том, что, хотя наши оценки по  $n$  выборкам (точкам сетки) могут быть шумными, они приближаются к некоторому надежному значению по мере того, как  $n \rightarrow \infty$ :

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n f(\Theta_i) \tilde{\mathcal{P}}(\Theta_i) \Delta \Theta_i}{\sum_{i=1}^n \tilde{\mathcal{P}}(\Theta_i) \Delta \Theta_i} = C \quad (28)$$

Последовательность - это впоследствии идея о том, что значение, к которому мы сходимся, является истинным значением, которое мы хотим оценить:

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n f(\Theta_i) \tilde{\mathcal{P}}(\Theta_i) \Delta \Theta_i}{\sum_{i=1}^n \tilde{\mathcal{P}}(\Theta_i) \Delta \Theta_i} = \mathbb{E}_{\mathcal{P}} [f(\Theta)] \quad (29)$$

Легко показать, что если матожидание хорошо определено (т.е. существует) и сетка покрывает всю область  $\Theta$  (т.е. охватывает наименьшие и наибольшие возможные значения в каждом измерении), то использование сетки является согласованным способом оценки матожидания. Это должно иметь интуитивный смысл: при условии, что наша сетка достаточно обширна в  $\Theta$ , так что мы не "пропустим" ни одной области пространства параметров, мы должны быть в состоянии оценить  $\mathbb{E}_{\mathcal{P}} [f(\Theta)]$  с произвольной точностью, просто увеличивая разрешение в  $\Delta \Theta$ .

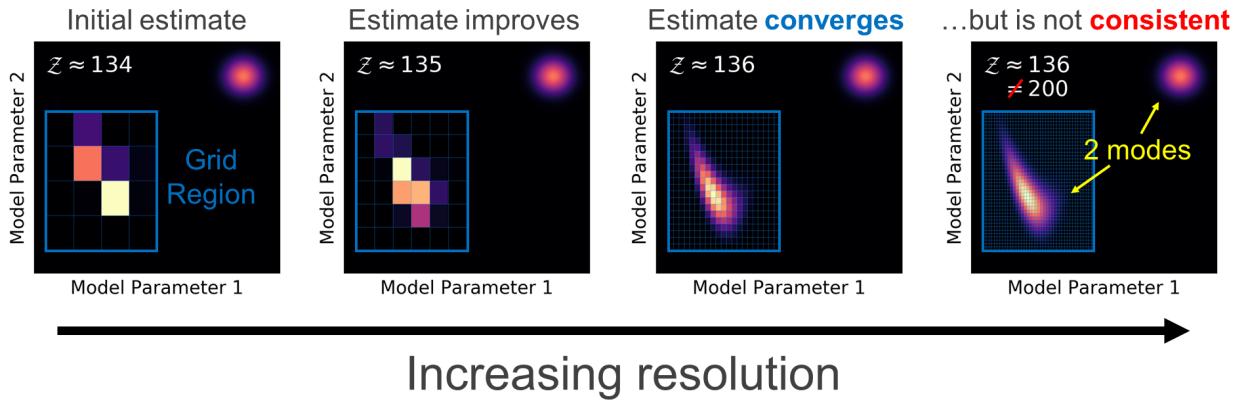


Рис. 5: Иллюстрация того, как оценки на основе сетки могут быть конвергентными (т.е. сходиться к одному значению по мере увеличения числа точек сетки), но не консистентными (т.е. значение, к которому они сходятся, не является правильным ответом). У нашего игрушечного двухмерного ненормированного апостериорного  $\tilde{P}(\Theta)$  есть два режима, которые хорошо разделены с общим доказательством  $Z = 200$ . Если мы не знаем о втором режиме, мы можем определить область сетки, которая охватывает только подмножество всего пространства параметров (слева). Хотя увеличение разрешения сетки в этой области позволяет оценкам  $Z$  сходиться к единому ответу (слева направо), это не равно правильному ответу  $Z = 200$ , потому что мы пренебрегли вкладом другой компоненты (справа). Дополнительные подробности см. в §4.3

К сожалению, мы не знаем заранее, в каком диапазоне значений  $\Theta$  должна находиться наша сетка. В то время как параметры могут лежать в диапазоне  $(-\infty, +\infty)$ , сетки опираются на элементы конечного объема, поэтому мы должны выбрать некоторое конечное подпространство для сетки. Поэтому, хотя сетки могут давать оценки, которые сходятся к некоторому значению в диапазоне, охватываемом точками сетки, всегда есть вероятность, что значительная часть апостериорного значения лежит за пределами этого диапазона. В таких случаях не гарантируется, что сетки будут последовательными оценщиками  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$ . Иллюстрация этой проблемы показана на [Figure 5](#). Этой фундаментальной проблеме не подвержены методы Монте-Карло, о которых я расскажу в §5.

**Упражнение:** Сетки над двумерным гауссовым пространством

### Setup

Рассмотрим ненормированный апостериор, хорошо аппроксимируемый двумерным гауссовым (нормальным) распределением с центром на  $(\mu_x, \mu_y)$  со стандартными отклонениями  $(\sigma_x, \sigma_y)$ :

$$\tilde{\mathcal{P}}(x, y) = \exp \left\{ -\frac{1}{2} \left[ \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} \right] \right\}$$

Предположим, что мы ожидаем найти, что наше апостериорное значение имеет среднее 0 и стандартное отклонение 1. Однако в действительности наше апостериорное значение имеет среднее  $(\mu_x, \mu_y) = (-0.3, 0.8)$  и стандартное отклонение  $(\sigma_x^2, \sigma_y^2) = (2.0, 5.0)$ , имитируя обычный случай, когда наши предварительные ожидания и апостериорные выводы несколько расходятся.

### Оценка на основе сетки

Мы хотим использовать двумерную сетку для оценки различных форм постреляционных интегралов. Начиная с равномерно расположенной сетки  $5 \times 5$  из  $[-2, 2]$ , вычислите:

1. доказательства  $\mathcal{Z}$ ,
2. средние  $\mathbb{E}_{\mathcal{P}}[x]$  и  $\mathbb{E}_{\mathcal{P}}[y]$ ,
3. 68% доверительные интервалы (или ближайшее приближение)  $[x_{\text{low}}, x_{\text{high}}]$  и  $[y_{\text{low}}, y_{\text{high}}]$ , элемент и эффективный размер выборки  $n_{\text{eff}}$ .

Насколько точно каждая из этих величин соответствует тем значениям, которые мы могли бы ожидать? Что говорит нам  $n_{\text{eff}}/n$  о том, насколько эффективно мы распределили точки сетки?

### Конвергенция

Повторите упражнение, используя равномерно расположенную сетку из  $20 \times 20$  точек и  $100 \times 100$  точек. Прокомментируйте любые различия. Насколько повысилась общая точность? Сходятся ли оценки?

## Консистенция

Далее расширьте границы сетки до  $[-5, 5]$  и выполните то же упражнение, что и выше. Существенно ли изменились ответы? Если да, то что это говорит нам о согласованности наших предыдущих оценок? Регулируйте плотность и границы сетки до тех пор, пока ответы не станут сходящимися и непротиворечивыми. Помните, что мы не знаем точной формы апостериорных оценок заранее. Что из этого следует в отношении общих проблем при применении сеток на практике?

## Эффективный размер выборки

Наконец, изучите, существует ли простая схема настройки расположения точек сетки  $x$  и  $y$  для максимизации эффективного размера выборки на основе определения, изложенного в §4.2. Если да, то можете ли вы объяснить, почему это работает? Если нет, то почему? Насколько адаптивная регулировка расстояния между сетками может улучшить  $n_{\text{eff}}$  и общую точность наших оценок по сравнению с эквивалентными равномерно распределенными сетками?

# 5 От сеток к методам Монте-Карло

## 5.1 Соединение точек сетки и выборок

Ранее я описал, как можно связать оценку  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$  с помощью сетки из  $n$  точек с эквивалентной оценкой с помощью набора  $n$  выборок  $\{f_1, \dots, f_n\}$  и ряда связанных с ними весов  $\{w_1, \dots, w_n\}$ . Основной результат состоит в том, что существует тесная связь между структурой апостериорной сетки и сетки с относительной амплитудой весов  $w_i \equiv \tilde{\mathcal{P}}(\Theta_i)\Delta\Theta_i$  для каждой точки  $f_i \equiv f(\Theta_i)$ . Регулировка разрешения сетки влияет на эти веса, при этом более равномерное распределение весов приводит к увеличению ESS, что может улучшить нашу оценку.

Тот факт, что уменьшение расстояния между точками (более плотная сетка) также уменьшает веса, имеет смысл: у нас больше точек, расположенных в этой области, поэтому каждая точка должна получить меньший относительный вес при вычислении  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$ . Аналогично, если у нас одинаковые расстояния между точками, но меняется относительная форма заднего плана, то вес этой точки при оценке  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$  также должен измениться соответственно.

Теперь я хочу расширить это базовое соотношение. Теоретически, адаптивное увеличение разрешения нашей сетки позволяет нам лучше контролировать элементы объема  $\Delta\Theta_i$ , используемые для получения весов. Если мы достаточно хорошо знаем форму нашего постера, то для больших  $n$  мы теоретически должны быть в состоянии настроить  $\Delta\Theta_i$  так, чтобы веса  $w_i = \tilde{\mathcal{P}}(\Theta_i)\Delta\Theta_i$  были равномерными с некоторой желаемой точностью. По проверке, это должно произойти, когда

$$\Delta\Theta_i \propto \frac{1}{\tilde{\mathcal{P}}(\Theta_i)} \quad (30)$$

для всех  $i$ .

Если довести эти рассуждения до концептуального предела, то с ростом  $n \rightarrow \infty$  мы можем представить, что оцениваем апостериор, используя все большее и большее количество точек сетки, расстояние между которыми  $\Delta\Theta$  меняется как функция от  $\Theta$ . Используя это, мы можем определить плотность точек  $\mathcal{Q}(\Theta)$  на основе изменяющегося разрешения  $\Delta\Theta(\Theta)$  нашей бесконечно тонкой сетки как функцию  $\Theta$ :

$$\mathcal{Q}(\Theta) \propto \frac{1}{\Delta\Theta(\Theta)} \quad (31)$$

Этот результат говорит о том, что в континуальном пределе, где  $n \rightarrow \infty$ , структура нашей сетки с бесконечным разрешением эквивалентна новому непрерывному распределению  $\mathcal{Q}(\Theta)$ . Иллюстрация этой концепции показана на [Figure 6](#). Используя  $\mathcal{Q}(\Theta)$ , мы можем переписать наше исходное матожидание в виде

$$\mathbb{E}_{\mathcal{P}}[f(\Theta)] \equiv \frac{\int f(\Theta) \tilde{\mathcal{P}}(\Theta) d\Theta}{\int \tilde{\mathcal{P}}(\Theta) d\Theta} = \frac{\int f(\Theta) \frac{\tilde{\mathcal{P}}(\Theta)}{\mathcal{Q}(\Theta)} \mathcal{Q}(\Theta) d\Theta}{\int \frac{\tilde{\mathcal{P}}(\Theta)}{\mathcal{Q}(\Theta)} \mathcal{Q}(\Theta) d\Theta} = \frac{\mathbb{E}_{\mathcal{Q}}[f(\Theta) \tilde{\mathcal{P}}(\Theta)/\mathcal{Q}(\Theta)]}{\mathbb{E}_{\mathcal{Q}}[\tilde{\mathcal{P}}(\Theta)/\mathcal{Q}(\Theta)]} \quad (32)$$

По причинам, которые скоро станут понятны, я буду называть  $\mathcal{Q}(\Theta)$  распределением предложений.

Сейчас это может показаться математическим трюком: все, что я сделал, это переписал наше исходное одно матожидание относительно (ненормированного) апостериорного  $\tilde{\mathcal{P}}(\Theta)$  в терминах два матожидания относительно распределения предложения  $\mathcal{Q}(\Theta)$ . Эта замена, однако, на самом деле позволяет нам полностью реализовать связь между точками сетки и выборками.

Ранее я показал, что оценка матожидания для точек сетки в точности аналогична оценке, которую мы получили бы, если бы точки сетки были случайными выборками  $\{f_1, \dots, f_n\}$  с соответствующими весами  $\{w_1, \dots, w_n\}$ . Однако после того, как мы определили наше ожидание относительно  $\mathcal{Q}(\Theta)$ , это утверждение может стать точным, если мы можем явно генерировать выборки из  $\mathcal{Q}(\Theta)$ .

Давайте быстро рассмотрим, что это значит. Изначально мы рассматривали попытку оценить  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$  по сетке с  $n$  точками. Однако в пределе бесконечного разрешения наша сетка становится эквивалентной некоторому распределению  $\mathcal{Q}(\Theta)$ . Используя  $\mathcal{Q}(\Theta)$ , мы можем переписать наше исходное выражение в терминах двух ожиданий,  $\mathbb{E}_{\mathcal{Q}}[f(\Theta) \tilde{\mathcal{P}}(\Theta)/\mathcal{Q}(\Theta)]$  и  $\mathbb{E}_{\mathcal{Q}}[\tilde{\mathcal{P}}(\Theta)/\mathcal{Q}(\Theta)]$ , над  $\mathcal{Q}(\Theta)$  вместо  $\mathcal{P}(\Theta)$ . Это помогает нам, потому что теоретически мы можем оценить эти конечные выражения явно, используя серию из  $n$  случайно сгенерированных выборок из  $\mathcal{Q}(\Theta)$ . Из-за случайности, присущей этому подходу, его принято называть Монте-Карло подходом для оценки  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$  из-за исторических связей со случайностью и азартными играми.

На первый взгляд, это удивительное утверждение. Когда мы вычисляем интеграл от функции  $f(\Theta)$  на ограниченной сетке, мы знаем, что в нашем приближении есть некоторая ошибка, связанная с дискретизацией сетки. Эта ошибка полностью deterministic: учитывая количество точек сетки  $n$  и определенную плотность дискретизации  $\mathcal{Q}(\Theta) \propto 1/\Delta\Theta(\Theta)$ , мы каждый раз будем получать один и тот же результат (и ошибку) для  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$ .

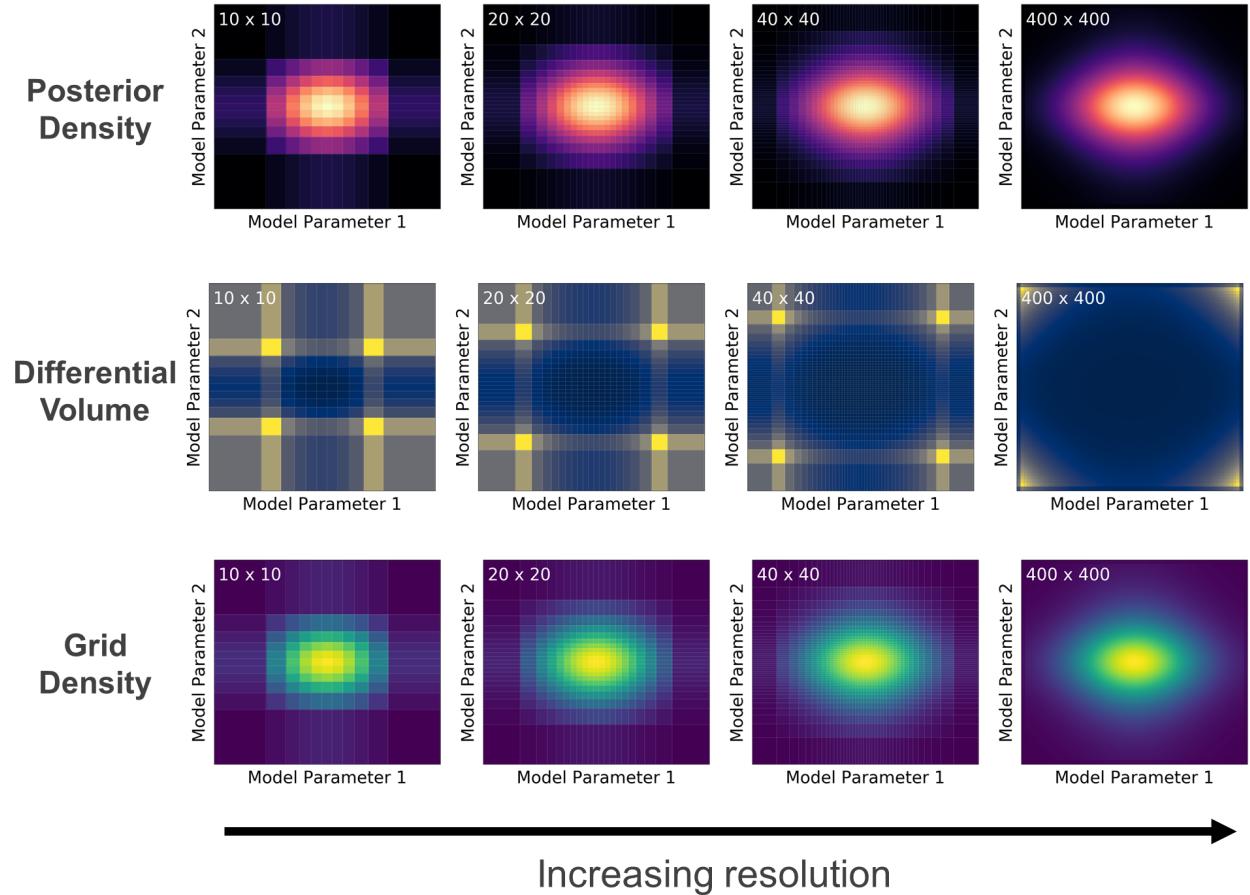


Рис. 6: Иллюстрация связи между сетками и непрерывными распределениями плотности. По мере увеличения числа точек сетки наша оценка апостериорного  $\mathcal{P}(\Theta)$  улучшается (вверху). Поскольку расстояние между точками сетки меняется, чтобы максимизировать эффективный размер выборки (см. Figure 4 и §4.2), элементы дифференциального объема  $\Delta\Theta_i$  меняются в зависимости от нашего местоположения (середина). При дальнейшем увеличении количества элементов объема плотность точек сетки в любом конкретном месте  $\rho(\Theta_i) = [\Delta\Theta_i]^{-1}$  ведет себя как непрерывная функция  $\mathcal{Q}(\Theta)$ , распределение которой похоже на  $\mathcal{P}(\Theta)$  (внизу). Это означает, что мы должны быть в состоянии использовать  $\mathcal{Q}(\Theta)$  каким-то образом для оценки  $\mathcal{P}(\Theta)$ . Дополнительные подробности см. в §5.

В отличие от этого, построение  $n$  образцов  $\{\Theta_1, \dots, \Theta_n\}$  из  $\mathcal{Q}(\Theta)$  является по своей сути рандомным (т.е. стохастическим) процессом, который не похож на сетку точек. И поскольку эти точки по своей природе случайны, фактическое отклонение между нашей оценкой и истинным значением  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$  будет тоже случайным. Таким образом, “ошибка” от случайных выборок говорит нам о том, насколько сильно может отличаться наша оценка от многих возможных реализаций нашего случайного процесса при определенном количестве выборок  $n$ , полученных из  $\mathcal{Q}(\Theta)$ . Тот факт, что мы можем получить примерно эквивалентные оценки на основе этих разных подходов при изменении  $n$  и  $\mathcal{Q}(\Theta)$ , лежит в основе связи между точками сетки и выборками.

Есть три основных преимущества перехода от адаптивно распределенной сетки к непрерывному распределению  $\mathcal{Q}(\Theta)$ . Во-первых, сетка всегда будет иметь некоторое минимальное разрешение  $\Delta\Theta_i$ , что затрудняет получение приблизительно равномерных весов, ограничивая максимальную ESS на практике. Напротив, теоретически мы можем добиться того, чтобы  $\mathcal{Q}(\Theta)$  более точно соответствовал апостериору  $\mathcal{P}(\Theta)$ , что даст большую ESS при фиксированном  $n$ .

Во-вторых, поскольку мы теперь работаем с распределениями, а не с конечным числом точек сетки, мы больше не ограничены некоторым конечным объемом при оценке ожиданий. Поскольку распределения могут лежать в диапазоне  $(-\infty, +\infty)$ , мы можем гарантировать, что  $\mathcal{Q}(\Theta)$  обеспечит достаточное покрытие по всем возможным значениям  $\Theta$ , по которым может быть определено наше послесловие  $\mathcal{P}(\Theta)$ . Это означает, что некоторые теоретические вопросы, поднятые в §4.3, связанные с применением сеток к апостериорам, которые варьируются в диапазоне  $(-\infty, +\infty)$ , больше не применимы. Таким образом, методы Монте-Карло могут служить согласованными оценками для более широкого диапазона возможных апостериорных ожиданий, чем методы на основе сеток, что делает их существенно более гибкими.

Наконец, минимальное число точек сетки всегда экспоненциально растет с размерностью (см. §4.1), независимо от того, сколько параметров нас интересует для маргинализации. Поскольку методы Монте-Карло не полагаются на них, они могут в полной мере использовать преимущества маргинализации по параметрам при оценке ожиданий  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$ . Поэтому они менее подвержены этому эффекту (хотя см. §7.2).

## 5.2 Выборка по важности

Как я уже пытался подчеркнуть ранее, основной постулат этой статьи заключается в том, что мы не знаем, как выглядит  $\mathcal{P}(\Theta)$  заранее. Это означает, что мы не знаем, какая структура сетки даст оптимальную оценку (т.е. максимум ESS) для  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$ , не говоря уже о том, как она должна вести себя в качестве  $\mathcal{Q}(\Theta)$  в континуальном пределе. Это дает нам достаточную мотивацию для того, чтобы выбрать  $\mathcal{Q}(\Theta)$  таким образом, чтобы сделать генерацию образцов из него простой и понятной.

Предположив, что мы выбрали такое  $\mathcal{Q}(\Theta)$ , мы можем впоследствии сгенерировать из него серию из  $n$  образцов. Предположим, что эти образцы имеют веса  $q_i$ , связанные с ними, и определим

$$f(\Theta_i) \equiv f_i, \quad \tilde{\mathcal{P}}(\Theta_i)/\mathcal{Q}(\Theta_i) \equiv \tilde{w}(\Theta_i) \equiv \tilde{w}_i \quad (33)$$

наше исходное выражение сводится к

$$\mathbb{E}_{\mathcal{P}} [f(\Theta)] = \frac{\mathbb{E}_{\mathcal{Q}} [f(\Theta)\tilde{w}(\Theta)]}{\mathbb{E}_{\mathcal{Q}} [\tilde{w}(\Theta)]} \approx \frac{\sum_{i=1}^n f_i \tilde{w}_i q_i}{\sum_{i=1}^n \tilde{w}_i q_i} \quad (34)$$

Если далее предположить, что мы выбрали  $\mathcal{Q}(\Theta)$  так, что можем моделировать выборки, которые независимо и идентично распределены (iid) (т.е. каждая выборка имеет то же распределение вероятности, что и другие, и все выборки взаимно независимы), то соответствующие веса выборок немедленно сводятся к  $q_i = 1/n$ , и наш результат становится

$$\mathbb{E}_{\mathcal{P}} [f(\Theta)] \approx \frac{n^{-1} \sum_{i=1}^n f_i \tilde{w}_i}{n^{-1} \sum_{i=1}^n \tilde{w}_i} \quad (35)$$

Как и в предыдущем случае с сетками (§4), знаменатель этого выражения снова является прямым приближением для доказательства

$$\mathcal{Z} = \int \tilde{\mathcal{P}}(\Theta) d\Theta \approx n^{-1} \sum_{i=1}^n \tilde{w}_i \quad (36)$$

Это дает прямой "рецепт" для оценки нашего исходного значения матожидания:

1. Извлеките  $n$  iid образцов  $\{\Theta_1, \dots, \Theta_n\}$  из  $\mathcal{Q}(\Theta)$ .
2. Вычислим их соответствующие веса  $\tilde{w}_i = \tilde{\mathcal{P}}(\Theta_i)/\mathcal{Q}(\Theta_i)$ .
3. Оценить  $\mathbb{E}_{\mathcal{P}} [f(\Theta)]$ . вычислив  $\mathbb{E}_{\mathcal{Q}} [\tilde{w}(\Theta)]$  и  $\mathbb{E}_{\mathcal{Q}} [f(\Theta)\tilde{w}(\Theta)]$  с использованием взвешенных выборочных средних.

Поскольку этот процесс заключается в "перевзвешивании" выборок на основе  $\tilde{w}_i$ , эти веса часто называют весами важности, а метод - Важной выборкой. Схематическая иллюстрация выборки по важности приведена на [Figure 7](#).

Мы можем интерпретировать веса важности как способ коррекции того, насколько "далека" наша первоначальная догадка  $\mathcal{Q}(\Theta)$  от истины  $\mathcal{P}(\Theta)$ . Если в позиции  $\Theta_i$  апостериорная плотность выше по сравнению с плотностью предложения, значит, мы с меньшей вероятностью сгенерировали выборку в этой позиции по сравнению с тем, что было бы, если бы мы брали выборки непосредственно из апостериорной. В результате мы должны увеличить соответствующий вес, чтобы учесть этот ожидаемый дефицит образцов в данной позиции. Если апостериорная плотность меньше плотности предложения, то верна альтернатива, и мы должны уменьшить вес соответствующего образца, чтобы учесть ожидаемый избыток образцов в данной позиции.

### 5.3 Примеры стратегий выборки

Выборка важности служит полезным первым шагом для понимания того, как веса  $\{\tilde{w}_1, \dots, \tilde{w}_n\}$  для соответствующего набора образцов  $n$  связаны с различными стратегиями выборки Монте-Карло.

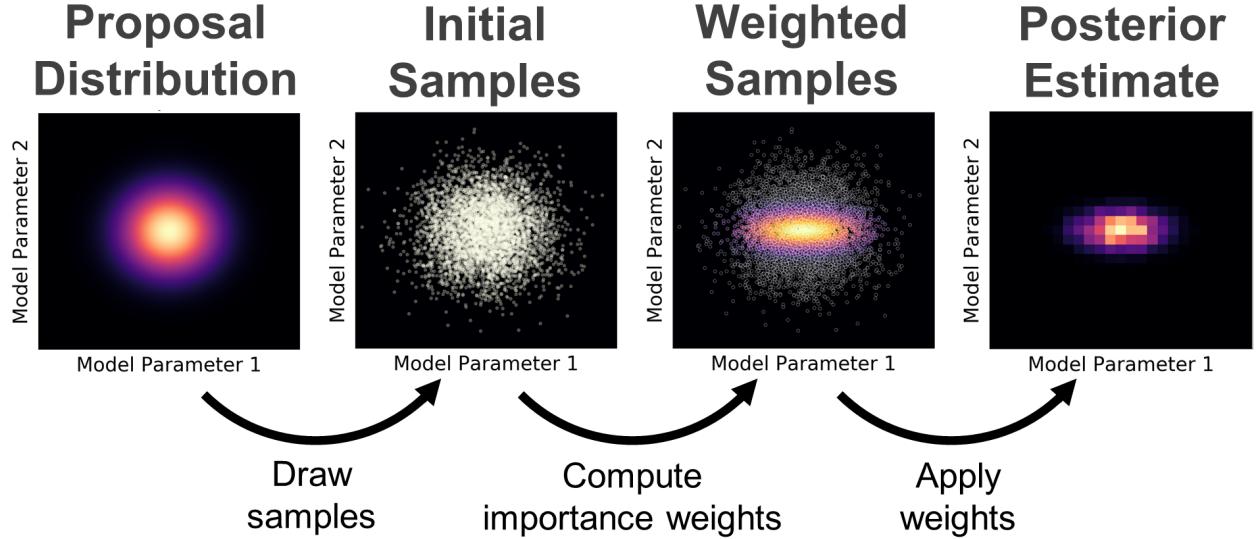


Рис. 7: Схематическая иллюстрация Importance Sampling. Сначала мы берем заданное распределение предложений  $Q(\Theta)$  (слева) и генерируем из него набор из  $n$  иидных выборок (в середине слева). Затем мы взвешиваем каждый образец на основе соответствующей “важности”  $\tilde{P}(\Theta)/Q(\Theta)$ , которую он имеет в данном месте (середина справа). Затем мы можем использовать эти взвешенные выборки для аппроксимации апостериорных ожиданий (справа). Дополнительные сведения см. в §5.2.

Например, одним из распространенных подходов является равномерная генерация образцов в пределах некоторого кубоида с объемом  $V$ . Тогда распределение предложения будет иметь вид

$$Q^{\text{unif}}(\Theta) = \begin{cases} 1/V & \Theta \text{ in cuboid} \\ 0 & \text{otherwise} \end{cases} \quad (37)$$

Соответствующие веса важности впоследствии будут просто пропорциональны апостериору в данной позиции:

$$\tilde{w}_i^{\text{unif}} = \frac{\tilde{P}(\Theta_i)}{Q^{\text{unif}}(\Theta_i)} = V\tilde{P}(\Theta_i) \propto P(\Theta_i) \quad (38)$$

Другой возможный подход заключается в том, чтобы принять наше предложение за предварительное:

$$Q^{\text{prior}}(\Theta) = \pi(\Theta) \quad (39)$$

Это кажется вполне оправданным выбором: предшествующее значение характеризует наши знания до изучения данных, поэтому оно должно служить полезной первой догадкой и охватывать диапазон всех возможностей. При таком предположении мы находим, что наши веса будут равны вероятности  $\mathcal{L}(\Theta)$  для каждой позиции:

$$w_i^{\text{prior}} = \frac{\tilde{P}(\Theta_i)}{Q^{\text{prior}}(\Theta_i)} = \frac{\mathcal{L}(\Theta_i)\pi(\Theta_i)}{\pi(\Theta_i)} = \mathcal{L}(\Theta_i) \quad (40)$$

Наконец, обратите внимание, что оптимальная стратегия выборки заключается в предположении, что мы можем принять наше предложение идентичным нашему апостериору:

$$\mathcal{Q}^{\text{post}}(\Theta) = \mathcal{P}(\Theta) \quad (41)$$

Тогда соответствующие веса будут просто постоянными и равными доказательству  $\mathcal{Z}$ :

$$w_i^{\text{post}} = \frac{\tilde{\mathcal{P}}(\Theta_i)}{\mathcal{Q}^{\text{post}}(\Theta_i)} = \frac{\mathcal{Z}\mathcal{P}(\Theta_i)}{\mathcal{P}(\Theta_i)} = \mathcal{Z} \quad (42)$$

Как и ожидалось, этот результат гарантирует максимально возможную ESS, равную  $n_{\text{eff}} = n$ . Таким образом, получение  $\mathcal{Q}(\Theta)$  как можно ближе к  $\mathcal{P}(\Theta)$  становится важной частью анализа при попытке использовать Importance Sampling для оценки значений ожиданий. Именно этот результат, в частности, мотивирует использование методов Марковской цепи Монте-Карло (MCMC), обсуждаемых с §6 и далее: если мы можем каким-то образом генерировать выборки прямо из  $\mathcal{P}(\Theta)$  или что-то близкое к нему, то мы можем достичь оптимальной оценки соответствующих значений ожиданий.

### Упражнение: Выборка по важности для двумерного гаусса

#### Setup

Вернемся к упражнению из §4, в котором наше ненормированное последействие хорошо аппроксимируется двумерным гауссовым (нормальным) распределением:

$$\tilde{\mathcal{P}}(x, y) = \exp \left\{ -\frac{1}{2} \left[ \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} \right] \right\}$$

где  $(\mu_x, \mu_y) = (-0.3, 0.8)$  и  $(\sigma_x^2, \sigma_y^2) = (2, 0.5)$ .

#### Выборка по значимости

Мы хотим использовать Importance Sampling для аппроксимации различных апостериорных интегралов от этого распределения. Начнем с того, что выберем распределение предложений  $\mathcal{Q}(x, y)$  как двумерное гауссовское со средним значением 0 и стандартным отклонением 1:

$$\mathcal{Q}(x, y) = \mathcal{N}[(\mu_x, \mu_y) = (0, 0), (\sigma_x, \sigma_y) = (1, 1)]$$

Используя  $n = 25$  iid случайных выборок, взятых из распределения предложения, вычислите оценку для:

1. доказательств  $\mathcal{Z}$ ,
2. средних  $\mathbb{E}_{\mathcal{P}}[x]$  и  $\mathbb{E}_{\mathcal{P}}[y]$ ,
3. 68% доверительные интервалы (или ближайшее приближение)  $[x_{\text{low}}, x_{\text{high}}]$  и  $[y_{\text{low}}, y_{\text{high}}]$ , элемент и эффективный размер выборки  $n_{\text{eff}}$ .

Насколько точно каждая из этих величин соответствует ожидаемым значениям? Что говорит нам  $n_{\text{eff}}/n$  о том, насколько хорошо наше предложение  $\mathcal{Q}(x, y)$  отслеживает базовое апостериорное значение  $\mathcal{P}(x, y)$ ?

## Неопределенность

Повторите это упражнение  $m = 100$  раз, чтобы получить оценку того, насколько сильно могут варьироваться наши оценки каждого количества. Соответствует ли эта вариация тому, что можно было бы ожидать, учитывая типичного эффективного размера выборки? Почему или почему нет?

## Сближение

Теперь повторите упражнение, используя точки  $n = 100$ ,  $n = 1000$  и  $n = 10000$ , а не  $n = 25$ , и прокомментируйте различия. Насколько повысилась общая точность? Сходятся ли и согласуются ли оценки при увеличении  $n_{\text{eff}}$ ? Насколько уменьшаются ошибки в оценках величин в зависимости от  $n$  и/или  $n_{\text{eff}}$ ? Ожидаемо ли такое поведение? Почему или почему нет?

## Последовательность

Далее расширим наше распределение предложений, чтобы вместо  $(\sigma_x, \sigma_y) = (2, 2)$  получить большее покрытие в “хвостах” апостериора. Выполните то же упражнение, что и выше, с  $n = \{100, 1000, 10000\}$  иидными случайными выборками. Существенно ли изменились ответы? Почему или почему нет?

Хотя теоретически мы можем выбрать  $\mathcal{Q}(x, y) \approx \mathcal{P}(x, y)$  так, чтобы  $n_{\text{eff}} \approx n$ , мы не знаем точной формы апостериора заранее. Учитывая, что  $\tilde{\mathcal{P}}(x, y)$  может отличаться от наших первоначальных ожиданий, что это упражнение говорит об общих проблемах применения Importance Sampling на практике?

## 6 Марковская цепь Монте-Карло

Теперь, когда мы видим, как веса связаны с различными стратегиями выборки Монте-Карло (например, генерацией выборок из предшествующих), я расскажу об идеи Марковской цепи Монте-Карло (МСМС). Вкратце, методы МСМС пытаются генерировать выборки таким образом, чтобы веса важности  $\{\tilde{w}_1, \dots, \tilde{w}_n\}$ , связанные с каждой выборкой, были постоянными. Основываясь на результатах из §5.3, это означает, что МСМС стремится генерировать выборки, пропорциональные апостериору  $\mathcal{P}(\Theta)$ , чтобы прийти к оптимальной оценке для нашего значения ожидания.

МСМС достигает этого, создавая цепочку (коррелированных) значений параметров  $\{\Theta_1 \rightarrow \dots \rightarrow \Theta_n\}$  за  $n$  итераций так, что число итераций  $m(\Theta_i)$ , проведенных в любой конкретной области  $\delta_{\Theta_i}$  с центром на  $\Theta_i$ , пропорционально плотности апостериора  $\mathcal{P}(\Theta_i)$ , содержащейся в этой области. Другими словами, “плотность” выборок, полученных в результате МСМС

$$\rho(\Theta) \equiv \frac{m(\Theta)}{n} \quad (43)$$

в позиции  $\Theta$ , интегрированной по  $\delta_{\Theta}$ , составляет приблизительно

$$\int_{\Theta \in \delta_{\Theta}} \mathcal{P}(\Theta) d\Theta \approx \int_{\Theta \in \delta_{\Theta}} \rho(\Theta) d\Theta \approx n^{-1} \sum_{j=1}^n \mathbf{1} [\Theta_j \in \delta_{\Theta}] \quad (44)$$

где  $\mathbb{1}[\cdot]$  - индикаторная функция, которая оценивается в 1, если внутреннее условие истинно, и в 0 в противном случае. Таким образом, мы можем аппроксимировать плотность простым сложением количества образцов в пределах  $\delta_{\Theta}$  и нормировкой на общее количество образцов  $n$ . Схематическая иллюстрация этой концепции показана на Figure 8.

Хотя это будет лишь приблизительно верно для любого конечного  $n$ , с ростом числа выборок  $n \rightarrow \infty$  эта процедура гарантирует, что  $\rho(\Theta) \rightarrow \mathcal{P}(\Theta)$  везде.<sup>2</sup> Теоретически, когда у нас есть достаточно разумное приближение для  $\rho(\Theta)$ , мы также можем использовать выборки  $\{\Theta_1 \rightarrow \dots \rightarrow \Theta_n\}$ , полученные из  $\rho(\Theta)$ , чтобы получить оценку для доказательства, используя тот же трюк с подстановкой, который был представлен в §5:

$$\mathcal{Z} = \int \frac{\tilde{\mathcal{P}}(\Theta)}{\rho(\Theta)} \rho(\Theta) d\Theta \equiv \mathbb{E}_{\rho} \left[ \tilde{\mathcal{P}}(\Theta)/\rho(\Theta) \right] \approx n^{-1} \sum_{i=1}^n \frac{\tilde{\mathcal{P}}(\Theta_i)}{\rho(\Theta_i)} \quad (45)$$

Это просто среднее значение отношения  $\tilde{\mathcal{P}}(\Theta_i)$  и  $\rho(\Theta_i)$  по всем  $n$  выборкам.

Наконец, поскольку наша процедура МCMC дает нам серию из  $n$  выборок из апостериорного значения, наше матожидание просто сводится к

$$\mathbb{E}_{\mathcal{P}} [f(\Theta)] \approx \frac{n^{-1} \sum_{i=1}^n f_i \tilde{w}_i}{n^{-1} \sum_{i=1}^n \tilde{w}_i} = \frac{n^{-1} \sum_{i=1}^n f_i}{n^{-1} \sum_{i=1}^n 1} = n^{-1} \sum_{i=1}^n f_i \quad (46)$$

Это просто выборочное среднее соответствующих  $\{f_1, \dots, f_n\}$  значений над нашим набором  $n$  выборок.

Я хотел бы остановиться на двух особенностях приведенных выше результатов, связанных с распространенными заблуждениями относительно методов МCMC. Во-первых, широко распространено мнение, что поскольку методы МCMC генерируют цепочку выборок, поведение которых следует за апостериором, мы не можем использовать их для оценки нормирующих констант, таких как доказательство  $\mathcal{Z}$ . Как было показано выше, это совсем не так: мы не только можем сделать это с помощью  $\rho(\Theta)$ , но и оценка, которую мы получаем, на самом деле является согласованной (хотя она будет сходиться медленно; см. §7.1).

Второе заблуждение заключается в том, что основной целью МCMC является “приближение” или “исследование” апостериорного значения. Другими словами, оценить  $\rho(\Theta)$ . Однако, как было показано выше, способность методов МCMC оценивать  $\rho(\Theta)$  действительно полезна только для оценки доказательств  $\mathcal{Z}$ . На самом деле, прослеживая наследие методов, основанных на выборке по важности, мы видим, что их основная цель - оценивать значения ожиданий (т.е. интегралы над апостериором). До этого момента я явно старался избегать упоминаний об “аппроксимации апостериорного значения”, чтобы избежать этого заблуждения, но я потрачу некоторое время на более подробное обсуждение этого момента в §7.1.

Вкратце, идея МCMC заключается в моделировании серии значений  $\{\Theta_1 \rightarrow \dots \rightarrow \Theta_n\}$  таким образом, чтобы их плотность  $\rho(\Theta)$  через заданный промежуток времени

<sup>2</sup>Обсуждение деталей того, когда/где именно это условие выполняется в теории и на практике, выходит за рамки данной статьи, но может быть найдено в других источниках, таких как Asmussen & Glynn (2011) и Brooks et al. (2011).

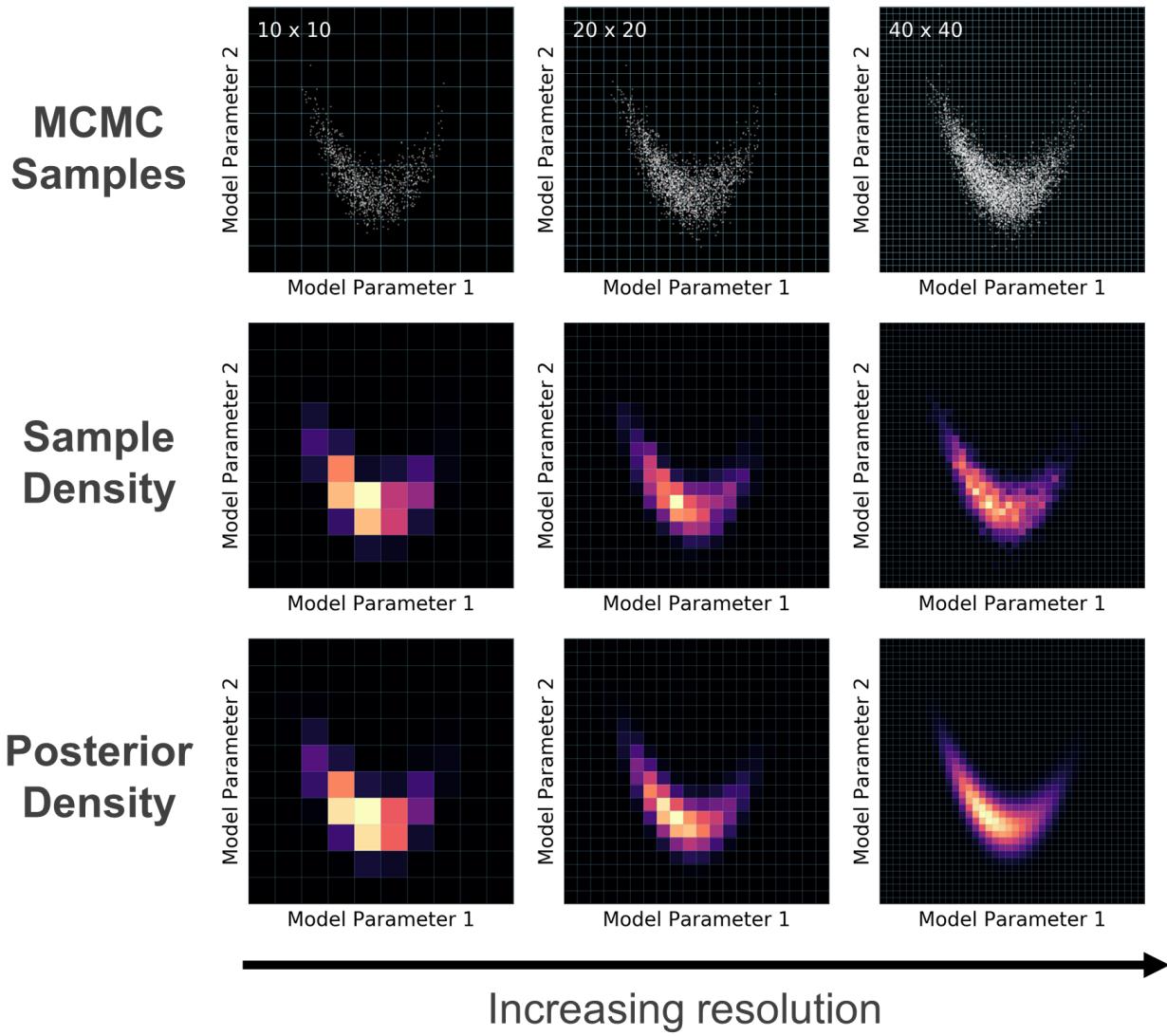


Рис. 8: Схематическая иллюстрация Марковской цепи Монте-Карло (ММС). ММС пытается создать цепочку из  $n$  (коррелированных) выборок  $\{\Theta_1 \rightarrow \dots \rightarrow \Theta_n\}$  (вверху) такую, что количество выборок  $m$  в некотором определенном объеме  $\delta$  дает относительную плотность  $m/n$  (в середине), сравнимую с апостериором  $\mathcal{P}(\Theta)$ , интегрированным по этому же объему (внизу). Дополнительные подробности см. в §6.

следовала базовому постеру  $\mathcal{P}(\Theta)$ . Затем мы можем оценить апостериор в любой конкретной области  $\delta_\Theta$ , просто подсчитав, сколько образцов мы там смоделировали, и нормировав на общее количество образцов  $n$ , которые мы сгенерировали. Поскольку мы также моделируем значения непосредственно из апостериорного ряда, любые матожидания также сводятся к простым выборочным средним. Эта процедура невероятно интуитивна и является одной из причин, по которой методы МСМС получили столь широкое распространение.

## 6.1 Генерирование выборок с помощью алгоритма Метрополиса-Гастингса

Существует обширная литература, посвященная различным подходам к генерации выборок (см., например, ссылки). Поскольку данная статья посвящена созданию концептуального понимания методов МСМС, изучение поведения большинства этих методов как в теории, так и на практике выходит за рамки данной статьи.

Вместо обзора я стремлюсь прояснить основы работы этих методов. Центральная идея заключается в том, что нам нужен способ генерировать новые выборки  $\Theta_i \rightarrow \Theta_{i+1}$  таким образом, чтобы распределение конечных выборок  $\rho(\Theta)$  по мере роста  $n \rightarrow \infty$  (1) было стационарным (то есть сходилось к чему-то) и (2) было равно  $\mathcal{P}(\Theta)$ . По сути, это аналоги ограничений сходимости и непротиворечивости, рассмотренных в §4.3.

Мы можем удовлетворить первое условие, вызвав detailed balance. Это идея о том, что вероятность сохраняется при переходе из одной позиции в другую (т.е. процесс обратим). Более формально это сводится к факторизации вероятности:

$$P(\Theta_{i+1}|\Theta_i)P(\Theta_i) = P(\Theta_{i+1}, \Theta_i) = P(\Theta_i|\Theta_{i+1})P(\Theta_{i+1}) \quad (47)$$

где  $P(\Theta_{i+1}|\Theta_i)$  - вероятность перехода из  $\Theta_i \rightarrow \Theta_{i+1}$ , а  $P(\Theta_i|\Theta_{i+1})$  - вероятность обратного перехода из  $\Theta_{i+1} \rightarrow \Theta_i$ . Перестановка дает следующее ограничение:

$$\frac{P(\Theta_{i+1}|\Theta_i)}{P(\Theta_i|\Theta_{i+1})} = \frac{P(\Theta_{i+1})}{P(\Theta_i)} = \frac{\mathcal{P}(\Theta_{i+1})}{\mathcal{P}(\Theta_i)} \quad (48)$$

где последнее равенство вытекает из того факта, что распределение, из которого мы пытаемся получить выборки, является апостериорным  $\mathcal{P}(\Theta)$ .

Теперь нам нужно реализовать процедуру, которая позволит нам фактически переходить на новые позиции, вычисляя эту вероятность. Мы можем сделать это, разбив каждое перемещение на два шага. Сначала мы хотим предложить новую позицию  $\Theta_i \rightarrow \Theta'_{i+1}$  на основе распределения предложений  $\mathcal{Q}(\Theta'_{i+1}|\Theta_i)$ , аналогичного  $\mathcal{Q}(\Theta)$ , используемого в Importance Sampling (§5.2). Затем мы либо примем решение принять новую позицию ( $\Theta_{i+1} = \Theta'_{i+1}$ ), либо отклонить новую позицию ( $\Theta_{i+1} = \Theta_i$ ) с некоторой вероятностью перехода  $T(\Theta'_{i+1}|\Theta_i)$ . Комбинируя эти условия, мы получаем вероятность перехода на новую позицию:

$$P(\Theta_{i+1}|\Theta_i) \equiv \mathcal{Q}(\Theta_{i+1}|\Theta_i)T(\Theta_{i+1}|\Theta_i) \quad (49)$$

Как и в случае с выборкой по важности, мы можем выбрать  $\mathcal{Q}(\Theta'_{i+1}|\Theta_i)$  таким образом, чтобы было просто предлагать новые образцы  $\Theta'_{i+1}$  путем численного моделирования. Затем нам нужно определить вероятность перехода  $T(\Theta'_{i+1}|\Theta_i)$  от того, должны

ли мы принять или отклонить  $\Theta'_{i+1}$ . Подставляя в наше выражение для детального баланса, мы обнаруживаем, что наша форма для вероятности перехода должна удовлетворять следующему ограничению:

$$\frac{T(\Theta_{i+1} | \Theta_i)}{T(\Theta_i | \Theta_{i+1})} = \frac{\mathcal{P}(\Theta_{i+1})}{\mathcal{P}(\Theta_i)} \frac{\mathcal{Q}(\Theta_i | \Theta_{i+1})}{\mathcal{Q}(\Theta_{i+1} | \Theta_i)} \quad (50)$$

Легко показать, что критерий Метрополиса Metropolis et al. (1953)

$$T(\Theta_{i+1} | \Theta_i) \equiv \min \left[ 1, \frac{\mathcal{P}(\Theta_{i+1})}{\mathcal{P}(\Theta_i)} \frac{\mathcal{Q}(\Theta_i | \Theta_{i+1})}{\mathcal{Q}(\Theta_{i+1} | \Theta_i)} \right] \quad (51)$$

удовлетворяет этому ограничению.

Генерировать образцы, следуя этому подходу, можно с помощью алгоритма Метрополиса-Гастингса (МН) (Metropolis et al., 1953; Hastings, 1970):

1. Предлагаем новую позицию  $\Theta_i \rightarrow \Theta'_{i+1}$ , генерируя выборку из распределения предложений  $\mathcal{Q}(\Theta'_{i+1} | \Theta_i)$ .
2. Вычислить вероятность перехода  $T(\Theta'_{i+1} | \Theta_i) = \min \left[ 1, \frac{\mathcal{P}(\Theta'_{i+1})}{\mathcal{P}(\Theta_i)} \frac{\mathcal{Q}(\Theta_i | \Theta'_{i+1})}{\mathcal{Q}(\Theta'_{i+1} | \Theta_i)} \right]$ .
3. Генерируем случайное число  $u_{i+1}$  из  $[0, 1]$ .
4. Если  $u_{i+1} \leq T(\Theta'_{i+1} | \Theta_i)$ , Принять ход и установите  $\Theta_{i+1} = \Theta'_{i+1}$ . Если  $u_{i+1} > T(\Theta'_{i+1} | \Theta_i)$ , отклонить ход и установите  $\Theta_{i+1} = \Theta_i$ .
5. Увеличиваем  $i = i + 1$  и повторяем этот процесс.

См. [Figure 9](#) для схематической иллюстрации этого процесса.

Поскольку алгоритмы, подобные алгоритму МН, генерируют цепочку состояний, в которых следующее предлагаемое положение зависит только от текущего положения, а не от любого из прошлых положений (т.е. он “забывает” прошлое), они известны как марковские процессы. Сочетание этих двух терминов с природой Монте-Карло для моделирования новых позиций и дало Марковской цепи Монте-Карло (МСМС) ее название.

Проблема с генерированием цепочки образцов на практике заключается в том, что наша цепочка имеет только конечную длину и начальную позицию  $\Theta_0$ . Если бы наша цепочка была бесконечно длинной, мы бы ожидали, что она посетит все возможные позиции в пространстве параметров, поэтому точная начальная позиция не имеет значения. Однако, поскольку на практике мы прекращаем выборку только после  $n$  итераций, старт с позиции  $\Theta_0$ , имеющей крайне низкую вероятность, означает, что непомерно большая доля наших  $n$  выборок займет эту низковероятную область, что может исказить наши окончательные результаты. Так как мы заранее имеем ограниченное представление о том, где находится  $\Theta_0$  относительно нашего апостериорного значения, на практике мы обычно хотим удалить начальную цепочку состояний, как только убедимся, что наша цепочка начала делать выборки из областей с более высокой вероятностью. Обсуждение различных подходов к определению и удалению образцов из этого периода сгорания выходит за рамки данной статьи; за дополнительной информацией обращайтесь к Gelman & Rubin (1992), Gelman et al. (2013), и Vehtari et al. (2019), а также к ссылкам на них.

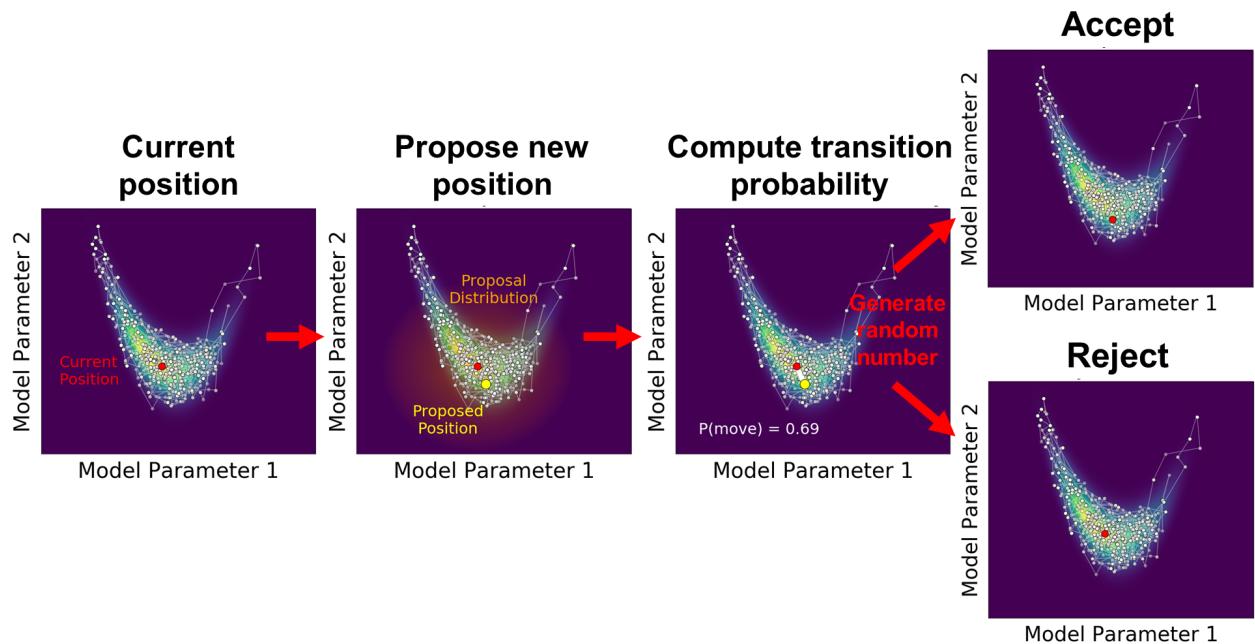


Рис. 9: Схематическая иллюстрация алгоритма Метрополиса-Хастингса. На заданной итерации  $i$  мы сгенерирували цепочку выборок  $\{\Theta_1 \rightarrow \dots \rightarrow \Theta_i\}$  (белые) до текущей позиции  $\Theta_i$  (красные), поведение которой следует за базовым постернным  $\mathcal{P}(\Theta)$  (цветовая карта Виридиса). Затем мы предлагаем новую позицию  $\Theta'_{i+1}$  (желтая) из распределения предложений (оранжевая заштрихованная область). Затем мы вычисляем вероятность перехода  $T(\Theta'_{i+1} | \Theta_i)$  (белый цвет) на основе апостериорных плотностей  $\mathcal{Q}(\Theta)$  и  $\mathcal{Q}(\Theta' | \Theta)$ . Затем мы генерируем случайное число  $u_{i+1}$  равномерно от 0 до 1. Если  $u_{i+1} \leq T(\Theta'_{i+1} | \Theta_i)$ , мы принимаем ход и делаем следующую позицию в цепочке  $\Theta_{i+1} = \Theta'_{i+1}$ . Если мы отклоняем ход, то  $\Theta_{i+1} = \Theta_i$ . Дополнительные подробности см. в §6.1.

## 6.2 Эффективный размер выборки и время автокорреляции

На данный момент кажется, что МСМС должен быть оптимальным методом для любой ситуации: моделируя выборки непосредственно из (неизвестного) заднего плана, мы можем получить оптимальную оценку для любого значения ожидания, которое хотим оценить. На практике, однако, это не так. Значения МСМС опираются на специальные алгоритмические процедуры, такие как алгоритм МН, для генерации выборок, чье предельное поведение сводится к цепочке выборок  $\{\Theta_1 \rightarrow \dots \rightarrow \Theta_n\}$ , чье распределение следует за апостериором. Однако любая заданная выборка  $\Theta_i$  с большей вероятностью будет коррелировать как с предыдущей выборкой в последовательности  $\Theta_{i-1}$ , так и с последующей выборкой в последовательности  $\Theta_{i+1}$ .

Это происходит по двум причинам. Во-первых, новые позиции  $\Theta_i$ , взятые из  $\mathcal{Q}(\Theta_i | \Theta_{i-1})$ , по конструкции имеют тенденцию зависеть от текущей позиции  $\Theta_{i-1}$ . Это означает, что позиция, которую мы предложим на итерации  $i + 1$ , будет коррелировать с позицией на итерации  $i$ , которая сама будет коррелировать с позицией на итерации  $i - 1$ , и т. д.

Во-вторых, даже если мы зададим  $\mathcal{Q}(\Theta' | \Theta) = \mathcal{Q}(\Theta')$ , чтобы все наши предложеные позиции были некоррелированы, наша вероятность перехода  $T(\Theta' | \Theta)$  все равно гарантирует, что мы в конечном итоге отклоним новую позицию так, что  $\Theta_{i+1} = \Theta_i$ . Поскольку образцы, находящиеся в одной и той же позиции, максимально коррелированы, это гарантирует, что образцы из нашей цепочки будут “в среднем” иметь ненулевые корреляции. Заметим, что при низких долях принятия (т.е. долях предложений, которые принимаются, а не отклоняются) большая часть цепочки будет содержать эти идеально коррелированные образцы, что увеличит общую корреляцию.

Как уже говорилось в §4.2, коррелированные выборки предоставляют меньше информации о базовом распределении, из которого они взяты, поскольку их поведение зависит не только от базового распределения, но и от соседних выборок в последовательности. Таким образом, выборки с более высокой степенью корреляции должны приводить к уменьшению ESS.

Эту интуицию можно выразить количественно, введя автоковариацию  $C(t)$  для некоторого целочисленного запаздывания  $t$ . Предположим, что у нас есть бесконечно длинная цепочка  $\{\Theta_1 \rightarrow \dots\}$ , автоковариация  $C(t)$  имеет вид:

$$C(t) \equiv \mathbb{E}_i [(\Theta_i - \bar{\Theta}) \cdot (\Theta_{i+t} - \bar{\Theta})] = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\Theta_i - \bar{\Theta}) \cdot (\Theta_{i+t} - \bar{\Theta}) \quad (52)$$

где  $\cdot$  - точечное произведение. Другими словами, мы хотим знать ковариацию между  $\Theta_i$  на некоторой итерации  $i$  и  $\Theta_{i+t}$  на некоторой другой итерации  $i+t$ , усредненную по всем возможным парам образцов  $(\Theta_i, \Theta_{i+t})$  в нашей бесконечно длинной цепочке. Заметим, что амплитуда  $|C(t)|$  будет максимальна при  $|C(t=0)|$ , когда два сравниваемых образца идентичны, и минимальна при  $|C(t)| = 0$ , когда  $\Theta_i$  и  $\Theta_{i+t}$  полностью независимы друг от друга.

Используя автоковариацию, мы можем определить соответствующую автокорреляцию  $A(t)$  как

$$A(t) \equiv \frac{C(t)}{C(0)} \quad (53)$$

Теперь измеряется средняя степень корреляции между выборками, разделенными цепочисленным лагом  $t$ . В случае, когда  $t = 0$ , обе выборки идентичны и  $A(t = 0) = 1$ . В случае, когда выборки некоррелированы на протяжении лага  $t$ ,  $A(t) = 0$ .

Общее время автокорреляции для нашей цепочки - это просто автокорреляция  $A(t)$ , просуммированная по всем ненулевым лагам ( $t \neq 0$ ):

$$\tau \equiv \sum_{t=-\infty}^{\infty} A(t) - 1 = 2 \sum_{t=1}^{\infty} A(t) \quad (54)$$

где  $-1$  следует из того, что автокорреляция без запаздывания равна  $A(t = 0) = 1$  (т.е. каждая выборка идеально коррелирует сама с собой), а подстановка обусловлена тем, что  $A(t) = A(-t)$  в силу симметрии. Если  $\tau = 0$ , то для того, чтобы выборки стали некоррелированными, не требуется никакого времени, и выборки можно считать иидными. Если  $\tau > 0$ , то в среднем требуется  $\tau$  дополнительных итераций, чтобы образцы стали некоррелированными. Иллюстрация этого процесса показана на [Figure 10](#).

Учет времени автокорреляции напрямую приводит к модифицированному определению ESS:

$$n'_{\text{eff}} \equiv \frac{n_{\text{eff}}}{1 + \tau} \quad (55)$$

На практике мы не можем точно вычислить  $\tau$ , поскольку у нас нет бесконечного числа выборок и мы не знаем  $\mathcal{P}(\Theta)$ . Поэтому часто требуется получить оценку  $\hat{\tau}$  времени автокорреляции, используя имеющийся набор образцов  $n$ . Хотя обсуждение различных подходов, используемых для получения  $\hat{\tau}$ , выходит за рамки данной работы, за дополнительными подробностями обращайтесь к Brooks et al. (2011).

Тот факт, что методы МСМС подвержены неотрицательным автокорреляционным временем ( $\tau \geq 0$ ), но имеют оптимальные веса важности  $\tilde{w}_i = 1$ , дает ESS в размере

$$n'_{\text{eff,MCMC}} = \frac{n_{\text{eff,MCMC}}}{1 + \tau} = \frac{n}{1 + \tau} \leq n \quad (56)$$

Это означает, что не существует гарантии, что МСМС всегда является оптимальным выбором для достижения наибольшей ESS. В частности, методы Importance Sampling, которые могут генерировать полностью иидные выборки без времени автокорреляции ( $\tau = 0$ ), но с неоптимальными весами важности  $\tilde{w}_i$ , вместо этого имеют ESS в размере

$$n'_{\text{eff,IS}} = \frac{n_{\text{eff,IS}}}{1 + \tau} = n_{\text{eff,IS}} = \frac{(\sum_{i=1}^n \tilde{w}_i)^2}{\sum_{i=1}^n \tilde{w}_i^2} \leq n \quad (57)$$

которая может быть больше, чем  $n'_{\text{eff,MCMC}}$  при фиксированном  $n$ .

Учитывая приведенные выше результаты, теперь должно быть ясно, что центральная мотивация методов МСМС заключается в том, могут ли они генерировать цепочку выборок с временем автокорреляции, достаточно малым, чтобы превзойти Importance Sampling. Верно ли это, будет зависеть от апостериорного значения, подхода, используемого для генерации цепочки выборок (см. §6.1 и §8) и распределения предложений  $\mathcal{Q}(\Theta)$ , используемого для Выборки по значимости (см. §5.3).

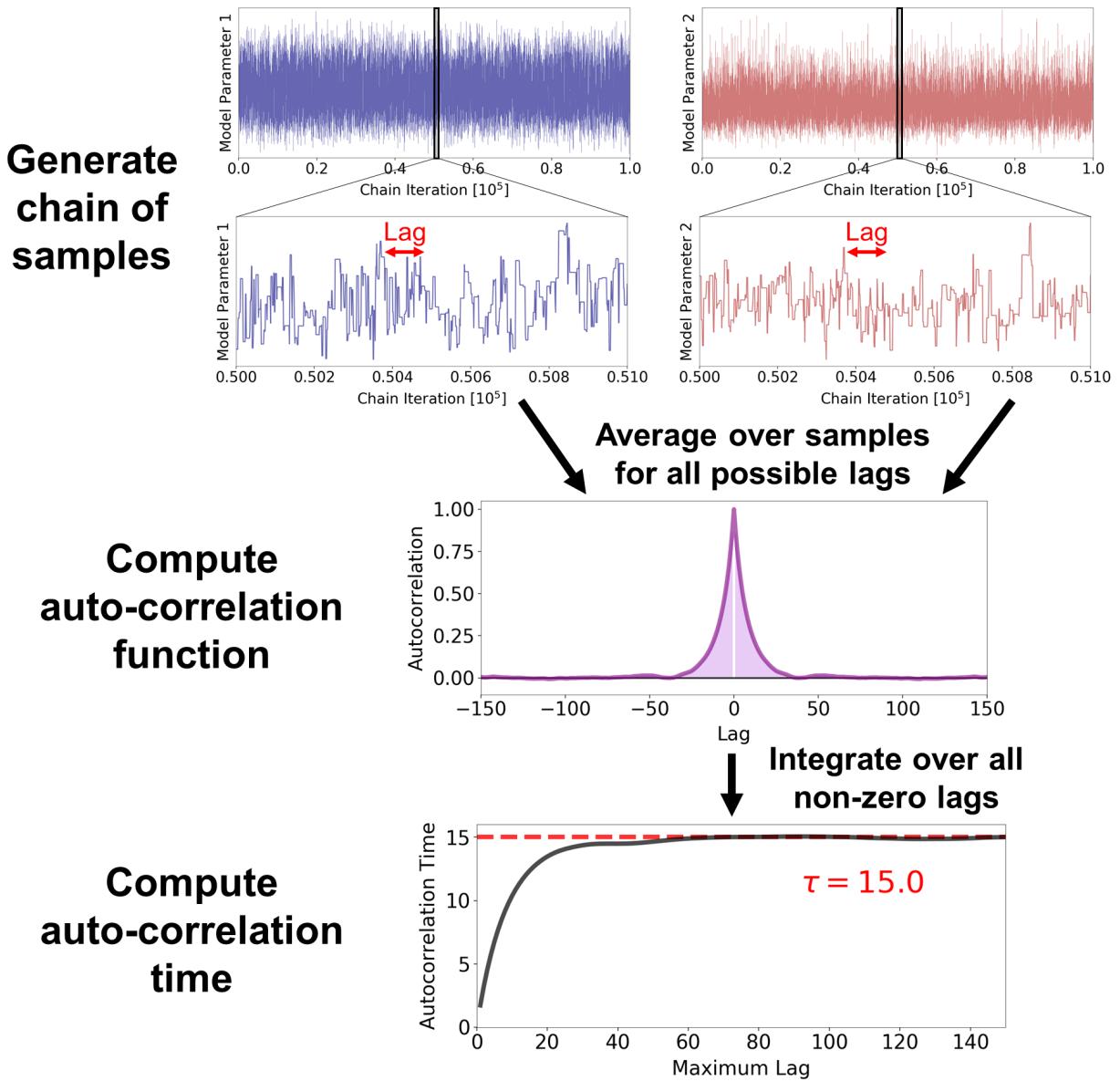


Рис. 10: Схематическая иллюстрация автокорреляции, связанной с МСМС. Методы МСМС генерируют цепочку выборок  $\{\Theta_1 \rightarrow \dots \rightarrow \Theta_n\}$  (вверху), но они имеют тенденцию быть сильно коррелированными на малых масштабах длины (вверху посередине). Мы можем количественно оценить степень корреляции, вычислив соответствующую автокорреляцию  $A(t)$  для нашего набора образцов и всех возможных временных задержек  $t$  (внизу посередине). Эта величина равна 1, когда  $t = 0$ , и уменьшается до 0, когда  $t \rightarrow \pm\infty$ . Общее время автокорреляции  $\tau$ , связанное с нашей цепочкой выборок, является просто интегрированной автокорреляцией по  $t \neq 0$ . Дополнительные подробности см. в §6.2.

## Exercise: MCMC over a 2-D Gaussian

### Setup

Let's again return to our examples from §4 and §5, in which our unnormalized posterior is well-approximated by a 2-D Gaussian (Normal) distribution:

$$\tilde{\mathcal{P}}(x, y) = \exp \left\{ -\frac{1}{2} \left[ \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} \right] \right\}$$

where  $(\mu_x, \mu_y) = (-0.3, 0.8)$  and  $(\sigma_x^2, \sigma_y^2) = (2, 0.5)$ .

We want to use MCMC to approximate various posterior integrals from this distribution. We will start by choosing our proposal distribution  $\mathcal{Q}(x', y' | x, y)$  to be a 2-D Gaussian with a mean of 0 and standard deviation of 1:

$$\mathcal{Q}(x', y' | x, y) = \mathcal{N}[(\mu_x, \mu_y) = (x, y), (\sigma_x, \sigma_y) = (1, 1)]$$

### Parameter Estimation

Using the above proposal, generate  $n = 1000$  samples following the MH algorithm starting from the position  $(x_0, y_0) = (0, 0)$ . Using these samples, compute an estimate of the means  $\mathbb{E}_{\mathcal{P}}[x]$  and  $\mathbb{E}_{\mathcal{P}}[y]$  as well as the corresponding 68% credible intervals (or closest approximation)  $[x_{\text{low}}, x_{\text{high}}]$  and  $[y_{\text{low}}, y_{\text{high}}]$ . How accurate are each of these quantities compared with the values we might expect?

### Evidence Estimation

Next, use a set of  $10 \times 10$  bins from  $x = [-5, 5]$  and  $y = [-5, 5]$  to construct an estimate  $\rho(x, y)$  from the resulting set of samples. Using this estimate for the density, compute an estimate of the evidence  $\mathcal{Z}$ . How accurate is our approximation? Does it substantially change if we adjust the number and/or size of the bins?

### Auto-Correlation Time and Effective Sample Size

Use numerical methods to compute an estimate of the auto-correlation time  $\tau$  and the corresponding effective sample size  $n_{\text{eff}}$ . How efficient is our sampling ( $n_{\text{eff}}/n$ ) compared to the default Importance Sampling approach from the exercise in §5? Does this mirror what we'd expect given the acceptance fraction of our proposals? What do these quantities tell us about how well our proposal  $\mathcal{Q}(x, y)$  matches the structure of the underlying posterior  $\mathcal{P}(x, y)$ ?

### Uncertainties

Repeat the above exercises  $m = 30$  times to get an estimate for how much our estimates of each quantity can vary. Is the variation in line with what might be expected given the typical effective sample size?

## Consistency and Convergence

Now repeat the above exercise using  $n = 2500$  and  $n = 10000$  samples points and comment on any differences. How much has the overall accuracy improved? Do the estimates appear convergent and consistent as  $n_{\text{eff}}$  increases? How much do the errors on quantities shrink as a function of  $n$  and/or  $n_{\text{eff}}$ ? Is this similar or different from the observed dependence from the Importance Sampling exercise in §5?

## Sampling Efficiency

Next, adjust the  $(\sigma_x, \sigma_y)$  of the proposal distribution to try and improve  $n_{\text{eff}}$  at fixed  $n$ . How close is the final ratio  $\sigma_x/\sigma_y$  of our proposal to that of the underlying posterior? Are there any additional scaling differences between the rough size of our proposal  $\mathcal{Q}(x', y'|x, y)$  relative to the underlying posterior  $\mathcal{P}(x, y)$ ? Given that  $\tilde{\mathcal{P}}(x, y)$  may differ from the structure assumed when picking  $\mathcal{Q}(x', y'|x, y)$ , can you think of any possible scheme to try and adjust our proposal using an existing set of samples?

## Burn-In

Finally, adjust the starting position to be at  $(x_0, y_0) = (10, 10)$  instead of  $(0, 0)$  and generate a new chain of samples. Plot the  $x$  and  $y$  positions of the chain over time. Are there any obvious signs of the burn-in period? How many samples roughly should be assigned to burn-in and subsequently removed from our chain? Are there any possible heuristics that might help to identify the initial burn-in period?

# 7 Выборка последователей с помощью МСМС

Подход, с помощью которого методы МСМС способны генерировать цепочку выборок, сразу же наводит на мысль, что наша цепочка "исследует" апостериор. Хотя верно, что плотность выборок из цепочки  $\rho(\Theta) \rightarrow \mathcal{P}(\Theta)$  как  $n \rightarrow \infty$ , основной целью МСМС является оценка ожидаемых значений  $\mathbb{E}_{\mathcal{P}} [f(\Theta)]$ . Хотя это может показаться тонкой разницей, на самом деле это различие имеет решающее значение для понимания того, как алгоритмы МСМС (должны) вести себя на практике. Ниже мы обсудим это более подробно.

## 7.1 Аппроксимация апостериора

Хотя алгоритмы, такие как МН (§6.1), построены таким образом, что плотность цепочки выборок  $\rho(\Theta)$ , генерируемых МСМС, сходится к апостериору  $\mathcal{P}(\Theta)$  при  $n \rightarrow \infty$ , это не обязательно приводит к эффективному методу аппроксимации апостериора на практике. Другими словами,  $n$  должно быть очень большим, чтобы это ограничение выполнялось. Так сколько же образцов нам нужно, чтобы убедиться, что  $\rho(\Theta)$  является хорошим приближением к  $\mathcal{P}(\Theta)$ ?

Для начала нам нужно определить некоторую метрику для того, что такое "хорошее" приближение. Разумным вариантом может быть то, что мы хотели бы знать апостериор

в некоторой области  $\delta_{\Theta}$  с точностью  $\epsilon$  так, чтобы

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{1} [\Theta_i \in \delta_{\Theta}] - \int_{\delta_{\Theta}} \mathcal{P}(\Theta) d\Theta \right| \equiv |\hat{p}(\delta_{\Theta}) - p(\delta_{\Theta})| < \epsilon \quad (58)$$

где  $p(\delta_{\Theta})$  - полная вероятность, содержащаяся в пределах  $\delta_{\Theta}$ , а  $\hat{p}(\delta_{\Theta})$  - доля МСМС-цепочки выборок, содержащаяся в той же области. Хотя может показаться странным оценивать это только для одной области, я вскоре обобщу это, чтобы охватить все<sup>3</sup>.

В идеальном случае, когда наши выборки являются иидными и взяты из  $\mathcal{P}(\Theta)$ , каждая из наших выборок имеет вероятность  $p(\delta_{\Theta})$  оказаться в пределах  $\delta_{\Theta}$ . Вероятность того, что  $\hat{p}(\delta_{\Theta}) = m/n$ , следует биномиальному распределению:

$$P\left(\hat{p}(\delta_{\Theta}) = \frac{m}{n}\right) = \binom{n}{m} [p(\delta_{\Theta})]^m [1 - p(\delta_{\Theta})]^{n-m} \quad (59)$$

Другими словами, наши образцы оказываются внутри  $\delta_{\Theta}$  всего  $m$  раз с вероятностью  $p(\delta_{\Theta})$  и вне  $\delta_{\Theta}$  всего  $n-m$  раз с вероятностью  $1-p(\delta_{\Theta})$ . Дополнительный биномиальный коэффициент  $\binom{n}{m}$  для “ $n$  выбрать  $m$ ” учитывает все возможные уникальные случаи, когда  $m$  образцов могут оказаться внутри  $\delta_{\Theta}$  из общего объема выборки в  $n$ .

Это распределение имеет среднее значение  $p(\delta_{\Theta})$ , поэтому для любого конечного  $n$  мы ожидаем, что  $\hat{p}(\delta_{\Theta})$  будет несмещенным оценщиком  $p(\delta_{\Theta})$ :

$$\mathbb{E}[\hat{p}(\delta_{\Theta}) - p(\delta_{\Theta})] = p(\delta_{\Theta}) - p(\delta_{\Theta}) = 0 \quad (60)$$

Однако дисперсия зависит от размера выборки:

$$\mathbb{E}[|\hat{p}(\delta_{\Theta}) - p(\delta_{\Theta})|^2] = \frac{p(\delta_{\Theta})[1 - p(\delta_{\Theta})]}{n} \quad (61)$$

На практике мы можем ожидать, что время автокорреляции  $\tau > 0$  будет ненулевым. Это увеличит количество МСМС-выборок, которые нам нужно будет сгенерировать, чтобы быть уверенными в том, что наша оценка  $\hat{p}(\delta_{\Theta})$  является благополучной. Подставив коэффициент  $1 + \tau$  и подставив наше ожидание в ограничение точности, мы получим грубое ограничение на количество выборок  $n$ , которое нам потребуется как функция от  $\epsilon$ :

$$n \gtrsim \frac{p(\delta_{\Theta})[1 - p(\delta_{\Theta})]}{\epsilon^2/(1 + \tau)} \sim \frac{\hat{p}(\delta_{\Theta})[1 - \hat{p}(\delta_{\Theta})]}{\epsilon^2} \times (1 + \hat{\tau}) \quad (62)$$

Последняя замена  $p(\delta_{\Theta})$  и  $\tau$  их зашумленными оценками  $\hat{p}(\delta_{\Theta})$  и  $\hat{\tau}$  обусловлена тем, что на практике мы не знаем ни  $p(\delta_{\Theta})$ , ни  $\tau$  (и то, и другое требует полного знания постера). Поэтому мы вынуждены полагаться на оценки, полученные на основе набора выборок  $n$ .

Теперь рассмотрим этот результат более внимательно. Как и ожидалось, общее количество выборок пропорционально  $1 + \hat{\tau}$ : если для создания независимых выборок

<sup>3</sup>Технически процедура, описанная в этом разделе, работает только для конечных объемов. Однако основная интуиция работает даже при неограниченных параметрах, хотя доказательство этих результатов выходит за рамки данной работы

требуется больше времени, значит, нам нужно больше выборок, чтобы быть уверенными в том, что мы хорошо охарактеризовали апостериор в данной области. Мы также видим, что  $n \propto \epsilon^{-2}$ , так что если мы хотим уменьшить ошибку в  $x$  раз, нам нужно увеличить размер выборки в  $x^2$  раз.

Более интересно поведение числителя. Обратите внимание, что  $\hat{p}(\delta_{\Theta}) [1 - \hat{p}(\delta_{\Theta})]$  максимизируется для  $\hat{p}(\delta_{\Theta}) = 0.5$ , и поэтому наибольший размер выборки необходим, когда мы делим наше апостериор прямо пополам. Во всех остальных случаях необходимый размер выборки будет меньше, потому что за пределами или внутри интересующей нас области будет больше образцов, чью информацию мы можем использовать. Точное значение  $\hat{p}(\delta_{\Theta})$ , конечно, зависит как от апостериорного  $\mathcal{P}(\Theta)$ , так и от целевой области  $\delta_{\Theta}$ : размер выборки, необходимый для приближения апостериорного значения к некоторому  $\epsilon$  вблизи пика распределения (малая область, где  $\mathcal{P}(\Theta)$  велико), вероятно, будет отличаться от размера выборки, необходимого для точной оценки хвостов распределения (большая область, где  $\mathcal{P}(\Theta)$  мало).

Хотя приведенный выше аргумент работает, если мы хотим оценить апостериор только в одной области, “преобразование в апостериор” подразумевает, что мы хотим, чтобы  $\rho(\Theta)$  стало хорошим приближением к  $\mathcal{P}(\Theta)$  всюду. Мы можем обеспечить выполнение этого нового требования, разбив наше послесловие на  $m$  различных подобластей  $\{\delta_{\Theta_1}, \dots, \delta_{\Theta_m}\}$  и требуя, чтобы каждая подобласть была хорошо ограничена:

$$|\hat{p}(\delta_{\Theta_1}) - p(\delta_{\Theta_1})| < \epsilon_1 \quad \dots \quad |\hat{p}(\delta_{\Theta_m}) - p(\delta_{\Theta_m})| < \epsilon_m \quad (63)$$

Подставляя ожидаемые ошибки для каждого из этих ограничений дает нам приблизительное ограничение на количество выборок  $n_j$ , которые нам нужны для оценки апостериорного значения в каждой области  $\delta_{\Theta_j}$ :

$$n_j \gtrsim \frac{\hat{p}(\delta_{\Theta_j}) [1 - \hat{p}(\delta_{\Theta_j})]}{\epsilon_j^2} \times (1 + \hat{\tau}) \quad (64)$$

Таким образом, общее количество образцов, которые нам нужны, находится просто:

$$n \gtrsim \sum_{j=1}^m n_j \quad (65)$$

Такой подход к разбиению апостериора на подобласти концептуально схож с подходами на основе сетки, описанными в §4. Как таковой, он также подвержен тем же недостаткам: мы ожидаем, что количество регионов  $m$  будет увеличиваться экспоненциально с числом измерений  $d$ . Например, если бы мы просто хотели разделить наш апостериор на  $m$  orthants, то в итоге мы получили бы  $m = 2^d$  областей: 2 в одномерном (слева направо), 4 в двухмерном (верхний-левый, нижний-левый, верхний-правый, нижний-правый), 8 в трехмерном и т.д.

Этот эффект означает, что в общем случае количество выборок, необходимых для того, чтобы  $\rho(\Theta)$  было хорошим приближением к  $\mathcal{P}(\Theta)$  для некоторой заданной точности  $\epsilon$ , будет иметь вид

$$n \gtrsim k^d \quad (66)$$

где  $k$  - константа, зависящая от требований к точности. Это ставит аппроксимацию полного апостериорного значения в режим "проклятия размерности" (см. §4.1).<sup>4</sup>

Хотя многие практики говорят о том, что МСМС является эффективным методом для "приближения апостериорного значения", на практике он редко используется для непосредственного приближения  $\mathcal{P}(\Theta)$ . Как обсуждается в §3 и показано на Figure 2, почти все величины, о которых сообщается в литературе до, опираются не на приближения к полному  $d$ -мерному апостериору, а на приближения к маргинальным распределениям, которые всегда ограничены не более чем  $k \lesssim 3$  параметрами за один раз. Акт по оставшимся  $d - k$  параметрам помогает противостоять проклятию размерности, проиллюстрированному здесь. Хотя технически справедливо сказать, что МСМС может "исследовать" маргинальные  $k$ -D-апостериоры для определенных ограниченных наборов параметров, такой язык часто может привести к большим заблуждениям, чем к пониманию.

## 7.2 Апостериорный объем

Основные следствия, изложенные в §7.1, являются более общими, чем в конкретном случае, когда мы представляем себе разбиение апостериорной части на ортантные или другие области. По сути, вычисление любого ожидания по апостериору  $\mathbb{E}_{\mathcal{P}}[f(\Theta)]$  требует интегрирования по entire domain наших параметров  $\Theta$ . Поэтому мы хотим понять, как ведет себя объем этой области (т.е. сколько существует комбинаций параметров). Как только мы поймем, как он ведет себя, мы можем начать пытаться количественно оценить, как это влияет на наши оценки.

Для начала рассмотрим  $d$ -мерный гиперкуб ( $d$ -куб) с длиной стороны  $\ell$  во всех  $d$  измерениях. Его объем масштабируется как

$$V(\ell) = \prod_{i=1}^d \ell = \ell^d \quad (67)$$

Дифференциальный элемент объема между  $\ell$  и  $\ell + d\ell$  is

$$dV(\ell) = (d \times \ell^{d-1}) \times (d\ell) \propto \ell^{d-1} \quad (68)$$

Это экспоненциальное масштабирование с размерностью означает, что объем становится все более сконцентрированным в тонких оболочках, расположенных в областях, все более удаленных от центра  $d$ -куба. В качестве примера рассмотрим масштаб длины

$$\ell_{50} = 2^{-1/d}\ell \quad (69)$$

который делит  $d$ -куб на две равные по размеру области с 50% объема, содержащегося внутри  $\ell_{50}$  и 50% объема, содержащегося снаружи  $\ell_{50}$ . В одномерном случае это дает  $\ell_{50}/\ell = 0,5$ , как мы и ожидали. В двумерном случае это дает  $\ell_{50}/\ell \approx 0.7$ . В 3-D  $\ell_{50}/\ell \approx 0.8$ . В 7-D -  $\ell_{50}/\ell \approx 0.9$ . К моменту перехода к 15-D мы имеем  $\ell_{50}/\ell \approx 0.95$ , что означает,

---

<sup>4</sup>Прямое следствие этого результата состоит в том, что, хотя оценки доказательств, полученные с помощью МСМС, are согласованы, скорость сходимости к базовому значению будет происходить экспоненциально медленнее по мере увеличения  $d$ .

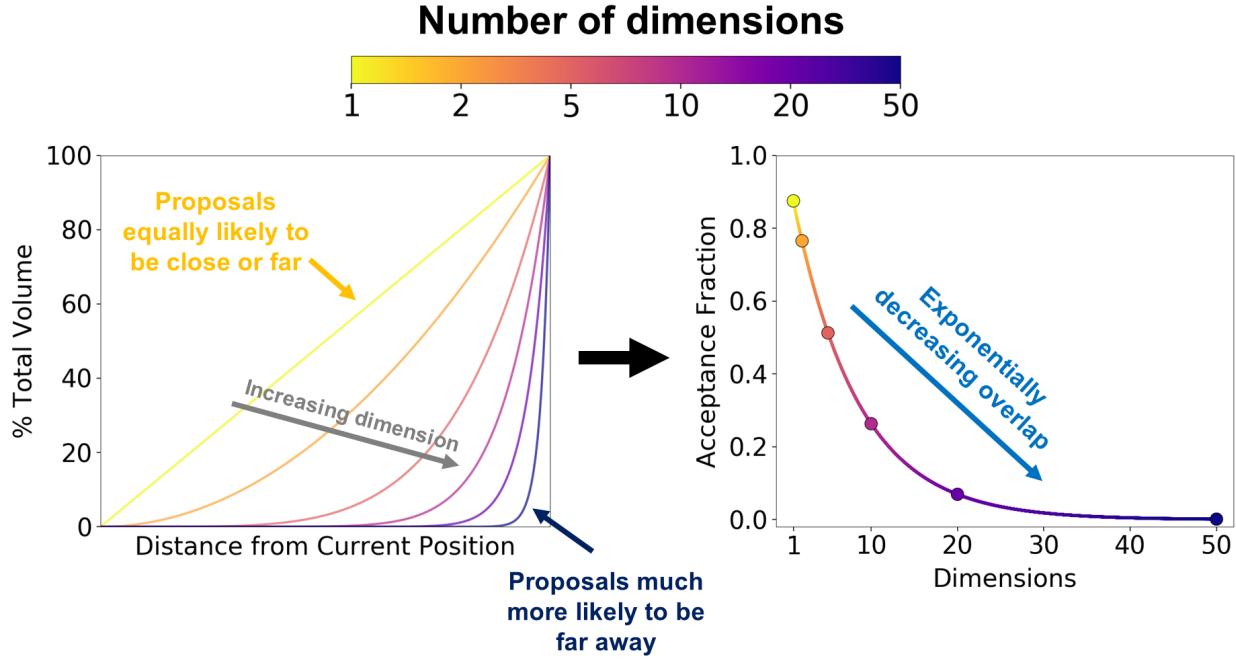


Рис. 11: Схематическая иллюстрация того, как проклятие размерности влияет на фракции принятия МСМС через объем апостериорных данных. При заданной позиции  $\Theta$  объем увеличивается  $\propto r^d$  как функция расстояния  $r$  от этой позиции (слева). По мере увеличения размерности этот объем становится все более концентрированным, что приводит к увеличению расстояния между предлагаемыми позициями  $\Theta'$  и текущей позицией  $\Theta$ . Большинство этих позиций имеют значительно меньшие апостериорные вероятности  $P(\Theta')$  по сравнению с текущим значением  $P(\Theta)$ , что приводит к экспоненциальному снижению типичной доли принятия (и соответствующему увеличению времени автокорреляции) по мере увеличения размерности (справа). Регулировка размера и/или формы предложения  $Q(\Theta'|\Theta)$  может помочь противостоять такому поведению. Дополнительные подробности см. в разделе §7.2.

что 50% объема находится в последних 5% шкалы длин вблизи границы  $d$ -куба. Хотя константы могут меняться при рассмотрении других форм (например, сферы), в целом это экспоненциальное масштабирование в зависимости от  $d$  является общим свойством высокоразмерных объемов. Другими словами, увеличение числа параметров приводит к экспоненциальному росту числа доступных комбинаций параметров, которые нам предстоит исследовать.

Помимо того, что этот экспоненциальный рост объема влияет на долгосрочное поведение МСМС, он также непосредственно влияет на то, как работают методы МСМС. Чтобы понять, почему это так, достаточно взглянуть на вероятности перехода, используемые в алгоритме МН, рассмотренном в §6.1:

$$T(\Theta_{i+1}|\Theta_i) \equiv \min \left[ 1, \frac{P(\Theta_{i+1}) Q(\Theta_i|\Theta_{i+1})}{P(\Theta_i) Q(\Theta_{i+1}|\Theta_i)} \right]$$

Нетривиальная часть этого выражения четко распадается на два члена. Первый зависит от *volume* и связан с тем, как мы предложили нашу следующую позицию из  $\mathcal{Q}(\Theta'|\Theta)$ . Вторая зависит от *density* и связана с тем, как меняется апостериорная плотность между двумя позициями.

На практике нашу вероятность перехода можно интерпретировать как базовый корректирующий подход: предложив новую позицию из некоторого близлежащего объема, мы затем пытаемся “скорректировать” различия между нашим предложением и базовым апостериором, принимая эти перемещения только иногда на основе изменений базовой плотности. В высоких измерениях это базовое “перетягивание каната” между объемом (предложение) и плотностью (апостериор) может разрушиться, поскольку большая часть объема объекта концентрируется у внешних краев.<sup>5</sup> Например, в случае, когда наше предложение  $\mathcal{Q}(\Theta'|\Theta)$  представляет собой куб с длиной стороны  $\ell$ , сосредоточенной на  $\Theta$ , это приводит к медианному масштабу длины  $\ell_{50} = 2^{-1/d}\ell$ , которая быстро увеличивается с  $0.5\ell$  до  $\approx \ell$  по мере увеличения размерности. Та же логика применима и к другим распределениям предложений (см. §8). Этот фокус на позициях, расположенных либо далеко, либо с очень похожими шкалами длины разделения, как  $\ell_{50} \rightarrow \ell$  означает, что многие варианты  $\mathcal{Q}(\Theta'|\Theta)$  имеют тенденцию к “проскачиванию”, предлагая новые позиции с гораздо меньшими плотностями апостериорного распределения по сравнению с текущей позицией. Эти новые позиции затем почти всегда отвергаются, что приводит к крайне низким долям принятия и соответственно большим временем автокорреляции. Пример этого эффекта показан на Figure 11.

Одним из основных способов борьбы с таким поведением является регулировка размера/формы предложения  $\mathcal{Q}(\Theta'|\Theta)$  таким образом, чтобы доля принятых предложений оставалась достаточно высокой. Это помогает гарантировать, что плотность постера  $\mathcal{P}(\Theta)$  не изменяется слишком сильно при предложении новых позиций, что приводит к снижению общего времени автокорреляции. Подробное описание того, как реализовать эти схемы на практике, выходит за рамки данной статьи; за дополнительными подробностями обращайтесь к citation.

### 7.3 Апостериорная масса и типичные сеты

Выше я описал, как поведение объема в высоких измерениях может повлиять на производительность нашего алгоритма выборки МСМС МН, возможно, приводя к неэффективным предложениям и низким долям принятия. Предположим, что мы решили эту проблему и имеем эффективный способ генерации нашей цепочки выборок. Теперь у нас есть вторичный вопрос: где находятся эти образцы?

Из нашего обсуждения в §7.1 мы знаем, что наибольшая плотность образцов  $\rho(\Theta)$  будет находиться там, где соответственно высока и апостериорная плотность  $\mathcal{P}(\Theta)$ . Однако эта область  $\delta_\Theta$  может соответствовать лишь небольшой части апостериорной плотности. Действительно, учитывая, что с ростом размерности объем экспоненциально увеличивается, почти гарантировано, что в моделях с большим количеством параметров  $\Theta$  подавляющая часть апостериорных данных будет находиться за пределами области

<sup>5</sup>Альтернативные методы, такие как гамильтоновский Монте-Карло,(Neal, 2012) могут обойти эту проблему, плавно включая изменения в плотности и объеме.

наибольшей плотности.

Следствием этого является то, что большинство образцов в нашей цепочке будут расположены вдали от пика плотности. В результате наша цепочка тратит большую часть своего времени на генерацию образцов в этих областях. Это оказывает огромное влияние на поведение нашей цепочки: в то время как наибольшая концентрация образцов будет располагаться в областях с наибольшей плотностью заднего плана, наибольшее количество образцов будет располагаться в областях с наибольшей массой заднего плана (т.е. плотность умножить на объем). Поскольку это подразумевает, что "типичный" образец (выбранный случайным образом), скорее всего, будет расположен в этой области с высокой задней массой, эту область также принято называть типичным множеством.

Чтобы немного упростить концепцию этого аргумента, представим, что у нас есть 3-параметрическая модель  $\Theta = (x, y, z)$  и  $\mathcal{P}(x, y, z)$  сферически симметрична. Хотя мы можем представить себе попытку интегрировать по  $\mathcal{P}(x, y, z)$  непосредственно в терминах  $dxdydz$ , почти всегда проще интегрировать по такому распределению в "оболочках" с дифференциальным объемом  $dV(r) = 4\pi r^2 dr$  как функцию радиуса  $r = \sqrt{x^2 + y^2 + z^2}$ . Это позволяет нам переписать трехмерный интеграл по  $(x, y, z)$  как одномерный интеграл по  $r$ :

$$\int \mathcal{P}(x, y, z) dxdydz = \int \mathcal{P}(r) 4\pi r^2 dr \equiv \int \mathcal{P}'(r) dr \quad (70)$$

где  $\mathcal{P}'(r) \equiv 4\pi r^2 \mathcal{P}(r)$  теперь является одномерной плотностью как функцией  $r$ . Это "увеличивает" вклад в функцию  $r$  дифференциального элемента объема оболочки, связанного с  $\mathcal{P}(r)$ , и подразумевает, что апостериор должен иметь какую-то оболочкоподобную структуру (т.е.  $\mathcal{P}'(r)$  максимизируется для  $r > 0$ ).

Хотя не все плотности апостериорных данных могут быть сферически-симметричными таким образом, в общем случае мы можем переписать  $d$ -D интеграл по  $\Theta$  как одномерный интеграл объема по  $V$ , определяемый некоторыми неизвестными изо-постериорными контурами<sup>6</sup>.

$$\int \mathcal{P}(\Theta) d\Theta = \int \mathcal{P}(V) dV \quad (71)$$

Как указано в §7.2, мы в общем случае ожидаем, что размер каждого элемента объема будет равен  $dV \sim r^{d-1} dr$ , где  $r$  - расстояние от пика постера. Таким образом, основная интуиция, которую мы получаем из простого сферически-симметричного случая, по-прежнему применима, и мы ожидаем, что

$$\int \mathcal{P}(V) dV \sim \int \mathcal{P}(r) r^{d-1} dr = \int \mathcal{P}'(r) dr \quad (72)$$

Как и прежде, дифференциальный элемент объема оболочки, связанный с  $\mathcal{P}(r)$ , "увеличивает" свой общий вклад как функция  $r$ . Это усиление также становится экспоненциально сильнее с ростом  $d$ . Для даже умеренно больших  $d$  мы ожидаем, что масса

---

<sup>6</sup>И действительно, альтернативные методы Монте-Карло, такие как Nested Sampling (Skilling, 2004, 2006) или Bridge/Path Sampling (Gelman & Meng, 1998), фактически предназначены для явной оценки такого типа интеграла объема

заднего плана будет в основном содержаться в тонкой оболочке, расположенной на радиусе  $r'$  с некоторой шириной  $\Delta r'$ . См. [Figure 12](#) для иллюстрации этого эффекта на основе игрушечной задачи, представленной в §8.1.

Этот результат имеет два непосредственных следствия. Во-первых, большинство наших выборок не находится там, где апостериорная плотность максимизируется. Это результат экспоненциально растущего числа комбинаций параметров, которые позволяют небольшой горстке отличных подгонок к данным легко быть подавленными значительно большим числом посредственных подгонок. Поэтому методы МСМС, как правило, неэффективны для определения местоположения и/или характеристики области пиковой апостериорной плотности.

Во-вторых, при увеличении  $d$  мы обычно ожидаем, что радиус оболочки, содержащей основную массу заднего плана, будет увеличиваться, удаляясь все дальше и дальше от пика плотности из-за экспоненциально увеличивающегося доступного объема. Поскольку большинство наших образцов находится в этой области, наша цепочка будет тратить подавляющее большинство времени на генерацию образцов из этой оболочки.

Это позволяет нам теперь точно описать, почему сложно эффективно предлагать образцы в высоких измерениях:

1. Чтобы наши доли принятия оставались разумными, нам нужно убедиться, что наши предложенные позиции в основном лежат внутри этой оболочки задней массы.
2. Однако получение независимой выборки требует возможности (теоретически) предложить любую позицию внутри этой оболочки.
3. Это означает, что время автокорреляции будет зависеть от того, сколько времени потребуется, чтобы "обойти" оболочку, что будет зависеть от ее общего размера  $r'$ , ширины  $\Delta r'$  и количества измерений  $d$ .

## 8 Применение к простой проблеме

Теперь я рассмотрю конкретный, подробный пример, чтобы проиллюстрировать, как все концепции, обсуждаемые в §6 и §7, объединяются на практике. На протяжении всего этого раздела я изложу ряд аналитических результатов и использую несколько различных стратегий МСМС-выборки для создания цепочек выборок. Я настоятельно рекомендую заинтересованным читателям реализовать свои собственные версии описанных здесь методов, которые могут быть использованы для полного воспроизведения численных результатов из этого раздела.

### 8.1 Toy Problem

In this toy problem, we will take our (unnormalized) posterior to be a  $d$ -dimensional Gaussian (Normal) distribution with a mean of  $\mu = 0$  and a standard deviation of  $\sigma$  in all dimensions:

$$\tilde{\mathcal{P}}(\Theta) = \exp \left[ -\frac{1}{2} \frac{|\Theta|^2}{\sigma^2} \right] \quad (73)$$

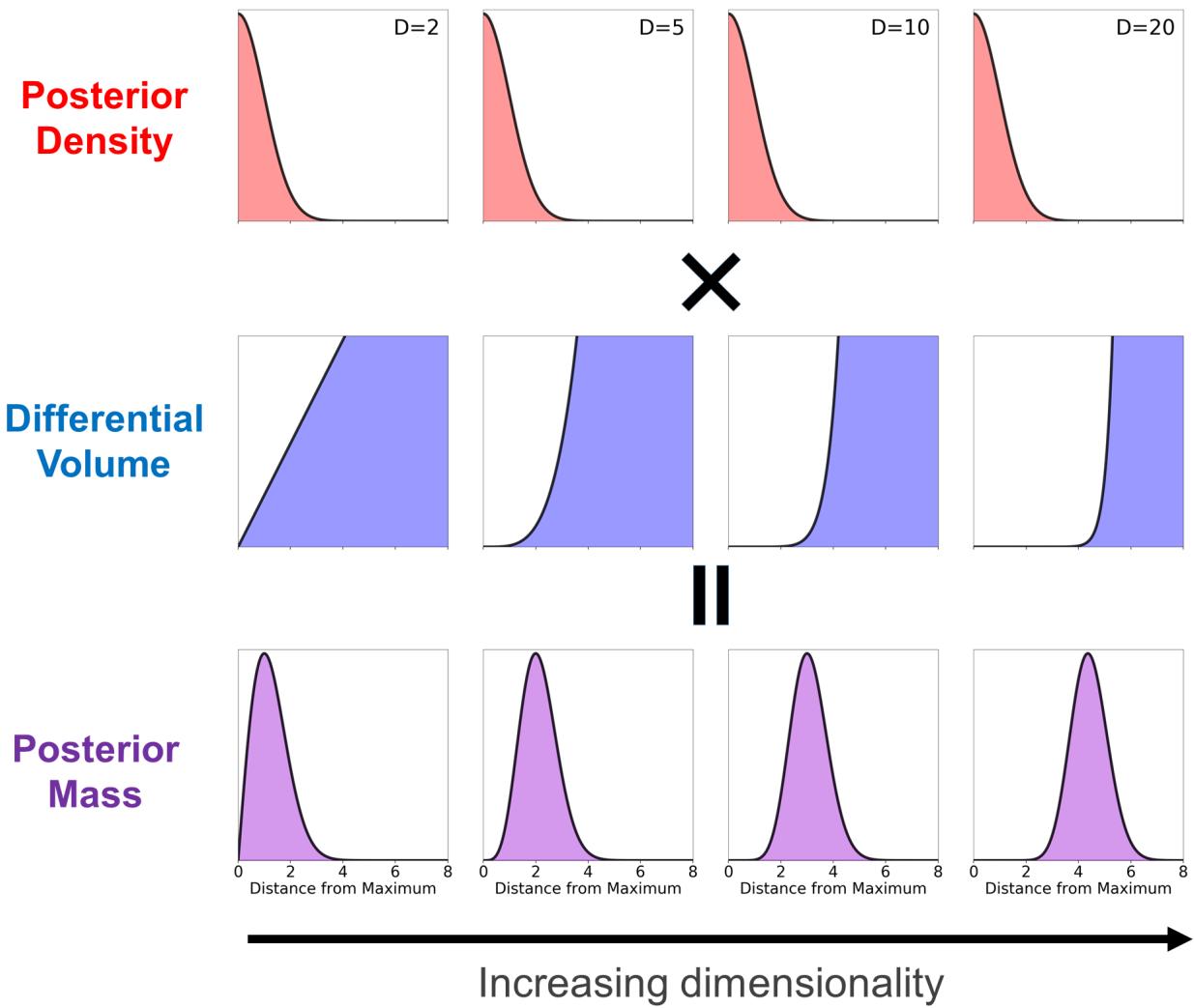


Рис. 12: Схематическая иллюстрация поведения апостериорной массы в зависимости от размерности с использованием  $d$ -мерного гаусса. На верхней панели показана апостериорная плотность  $\mathcal{P}(r) \propto e^{-r^2/2}$  (красный), построенная как функция расстояния  $r$  от максимальной апостериорной плотности при  $r = 0$  по мере увеличения числа измерений  $d$  (слева направо). Как и ожидалось, это распределение остается постоянным. На средней панели показан дифференциальный элемент объема  $dV(r) \propto r^{d-1} dr$  (синий) соответствующей оболочки радиусом  $r$ . Это иллюстрирует экспоненциально возрастающий объем, вносимый оболочками, удаленными от максимума. На нижней панели показана соответствующая “задняя масса” как функция радиуса  $\mathcal{P}'(r) \propto r^{d-1} \mathcal{P}(r) \propto r^{d-1} e^{-r^2/2}$  (фиолетовый). Из-за увеличения объема, расположенного дальше от максимума апостериорной плотности, мы видим, что большая часть апостериорной массы (и, следовательно, любых выборок, которые мы генерируем с помощью МСМС) на самом деле находится в оболочке, расположенной далеко от  $r = 0$ . Дополнительные подробности см. в §7.3.

where  $|\Theta|^2 = \sum_{i=1}^d \Theta_i^2$  is the squared magnitude of the position vector.

Based on the results from §7.3, we can better understand the properties of this distribution by rewriting the posterior density in terms of the “radius”  $r \equiv |\Theta| = \sqrt{\sum_{i=1}^d \Theta_i^2}$  away from the center:

$$\tilde{\mathcal{P}}(r) = \exp\left[-\frac{r^2}{2\sigma^2}\right] \quad (74)$$

The corresponding volume contained within a given radius  $r$  is then

$$V(r) \propto r^d \quad (75)$$

The corresponding posterior mass is  $\tilde{\mathcal{P}}'(r)$  is then defined via

$$\tilde{\mathcal{P}}(V)dV(r) \propto e^{-r^2/2\sigma^2} r^{d-1} dr \equiv \tilde{\mathcal{P}}'(r)dr$$

Note that this is closely related to the chi-square distribution.

The typical radius  $r_{\text{peak}}$  where the posterior mass peaks (i.e. is maximized) and a sample is most likely to be located can be derived by setting  $d\tilde{\mathcal{P}}'(r)/dr = 0$ . Solving this gives

$$r_{\text{peak}} = \sqrt{d-1}\sigma \quad (76)$$

In other words, while in 1-D a typical sample is most likely to be located at the peak of the distribution with  $r_{\text{peak}} = 0$ , in higher dimensions this changes quite drastically. While  $r_{\text{peak}} = 1\sigma$  in 2-D, it is  $2\sigma$  in 5-D,  $3\sigma$  in 10-D, and  $5\sigma$  in 26-D. This is a direct consequence of the huge amount of volume at larger radii in high dimensions: although a sample at  $r = 5\sigma$  has a posterior density  $\tilde{\mathcal{P}}(r)$  orders of magnitude worse than a sample at  $r = 0$ , the enormous number of parameter combinations (volume) available at  $r = 5\sigma$  more than makes up for it.

In general, we expect the posterior mass to comprise a “Gaussian shell” centered at some radius

$$r_{\text{mean}} \equiv \mathbb{E}_{\tilde{\mathcal{P}}'}[r] = \int_0^\infty r \tilde{\mathcal{P}}'(r) dr = \sqrt{2} \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} \sigma \approx \sqrt{d}\sigma \quad (77)$$

with a standard deviation of

$$\Delta r_{\text{mean}} \equiv \sqrt{\mathbb{E}_{\tilde{\mathcal{P}}'}[(r - r_{\text{mean}})^2]} = \sigma \sqrt{d - 2 \left( \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} \right)^2} \approx \frac{\sigma}{\sqrt{2}} \quad (78)$$

where  $\Gamma(d)$  is the Gamma function and the approximations are taken for large  $d$ . See [Figure 12](#) for an illustration of this behavior.

## 8.2 MCMC with Gaussian Proposals

Let us now consider a chain of samples  $\{\Theta_1 \rightarrow \dots \rightarrow \Theta_n\}$ . The distance between two samples  $\Theta_m$  and  $\Theta_{m+t}$  separated by some lag  $t$  will be

$$|\Theta - \Theta'| = \sqrt{\sum_{i=1}^d (\Theta_{m,i} - \Theta_{m+t,i})^2} \quad (79)$$

Assuming that the lag  $t \gg \tau$  is substantially larger than the auto-correlation time  $\tau$ , we can assume each sample is approximately iid distributed following our Gaussian posterior. This then gives an expected separation of

$$\Delta r_{\text{sep}} \equiv \sqrt{\mathbb{E}_{\mathcal{P}} [|\Theta_m - \Theta_{m+t}|^2]} = \sqrt{\sum_{i=1}^d \mathbb{E}_{\mathcal{P}} [(\Theta_{m,i} - \Theta_{m+t,i})^2]} = \sqrt{2d}\sigma \approx \sqrt{2}r_{\text{mean}} \quad (80)$$

We can in theory propose samples in such a way so that the separation  $|\Theta_{i+1} - \Theta_i|$  between a proposed position  $\Theta_{i+1}$  and the current position  $\Theta_i$  follows the ideal separation of  $\sqrt{2}r_{\text{mean}}$  derived above by using a simple Gaussian proposal distribution:

$$\mathcal{Q}(\Theta_{i+1} | \Theta_i) \propto \exp \left[ -\frac{1}{2} \frac{|\Theta_{i+1} - \Theta_i|^2}{2\sigma^2} \right] \quad (81)$$

While this proposal has the same shape as the posterior, it is centered on  $\Theta_i$  rather than 0. Using our intuition for how volume behaves based on §7.2, we can conclude that the majority of samples proposed from this choice of  $\mathcal{Q}(\Theta' | \Theta)$  will probably have little overlap with the posterior.

Indeed, numerical simulation suggests the typical fraction of positions that will be accepted given the above proposal roughly scales as

$$\langle f_{\text{acc}}(d) \rangle \equiv \exp [\mathbb{E}_{\mathcal{P}, \mathcal{Q}} [\ln T(\Theta_{i+1} | \Theta_i)]] \sim \exp \left[ -\frac{d}{4} - \frac{1}{2} \right] \quad (82)$$

which decreases exponentially as the dimensionality increases, similar to Figure 11. Likewise, we find the auto-correlation time roughly scales as

$$\langle \tau(d) \rangle \equiv \exp [\mathbb{E}_{\mathcal{P}, \mathcal{Q}} [\ln \tau]] \sim \exp \left[ \frac{d}{4} + \frac{7}{4} \right] \quad (83)$$

This exponential dependence arises because the overlap between the typical Gaussian proposal  $\mathcal{Q}(\Theta' | \Theta)$  and the underlying posterior  $\mathcal{P}(\Theta)$  essentially reduces to the small volume where two thin shells overlap. Since the radii of the shells goes as  $\propto \sqrt{d}$  while the widths remain roughly constant, the “fractional size” of the shell (and the corresponding overlap) ends up decreasing exponentially.

To counteract this effect, we need to adjust the  $\sigma$  of our proposal distribution by some factor  $\gamma$ :

$$\mathcal{Q}_\gamma(\Theta_{i+1} | \Theta_i) \propto \exp \left[ -\frac{1}{2} \frac{|\Theta_{i+1} - \Theta_i|^2}{(\gamma\sigma)^2} \right] \quad (84)$$

where our previous proposal assumes  $\gamma = \sqrt{2}$ . If we want to ensure our typical acceptance fraction will remain roughly constant as a function of dimension  $d$ ,  $\gamma$  needs to scale as

$$\langle f_{\text{acc}}(\gamma(d)) \rangle \approx C \Rightarrow \gamma(d) \propto \frac{1}{\sqrt{d}} \quad (85)$$

which inversely tracks the expected radius  $r_{\text{mean}}$  of the typical set. We find that taking

$$\gamma = \frac{\delta}{\sqrt{d}} \quad (86)$$

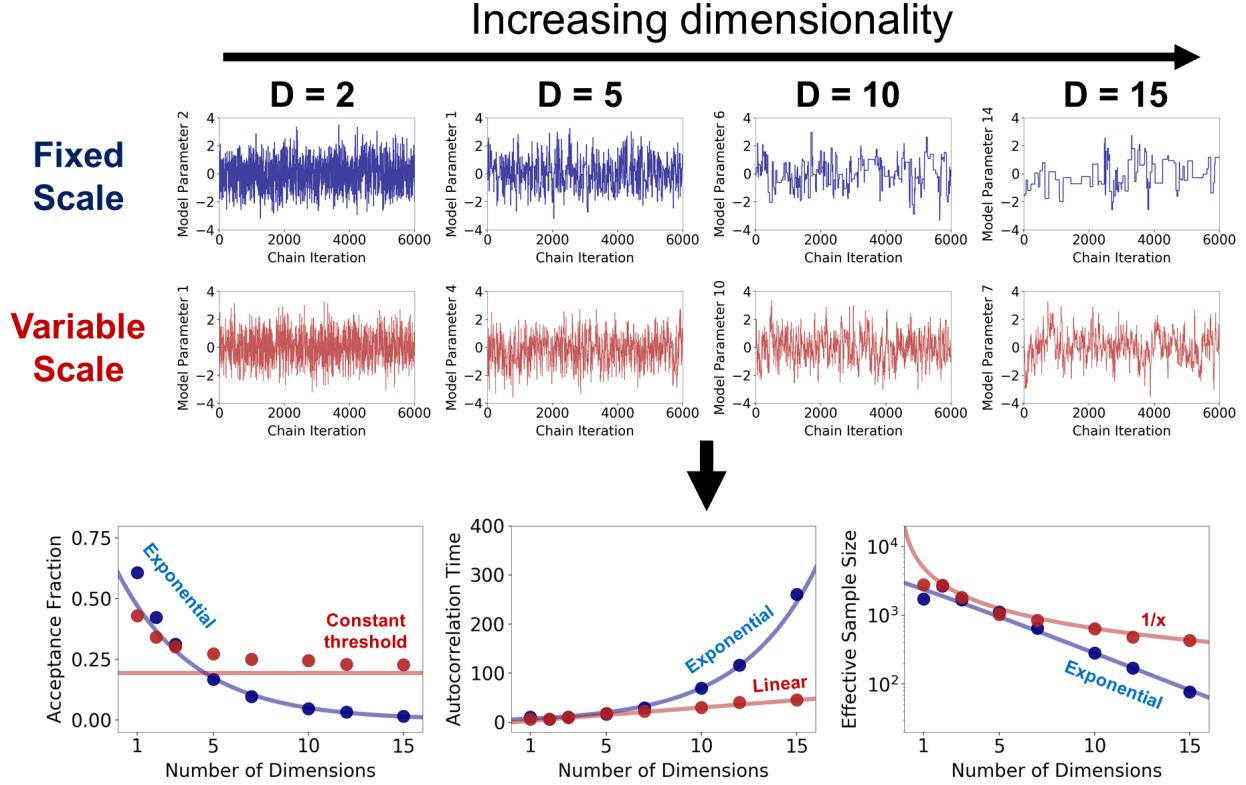


FIG. 13: Numerical results showcasing the performance of a simple MH MCMC sampler with Gaussian proposals on our toy problem, a  $d$ -dimensional Gaussian with mean  $\mu = 0$  and standard deviation  $\sigma = 1$  in every dimension. The top series of panels show snapshots of a random parameter from the chain as a function of dimensionality (increasing from left to right) assuming an unchanging proposal with constant scale factor  $\gamma = \sqrt{2}$  (blue) and a shrinking proposal with  $\gamma = 2.5/\sqrt{d}$  designed to target a constant acceptance fraction of  $\sim 25\%$  (red). The bottom panels show the corresponding acceptance fractions (left), auto-correlation times (middle), and effective sample sizes (right) from our chains (colored points) as a function of dimensionality. The approximations from §8.2 are shown as light colored lines. Shrinking the size of the proposal helps to keep samples within the bulk of the posterior mass, substantially reducing the auto-correlation time and increasing the effective sample size. Failing to do so leads to an exponentially decreasing fraction of good proposals and a corresponding exponential increase/decrease in the auto-correlation time/effective sample size. See §8.2 for additional discussion.

leads to a typical acceptance fraction of

$$\langle f_{\text{acc}}(\delta/\sqrt{d}) \rangle \approx \exp \left[ - \left( \frac{\delta^2}{4} \right)^2 - \frac{\delta}{2} \right] \quad (87)$$

as  $d$  becomes large with a typical auto-correlation time of

$$\langle \tau(\delta/\sqrt{d}) \rangle \approx 3d \quad (88)$$

for reasonable choices of  $\delta$ . This linear dependence is a substantial improvement over our earlier exponential scaling.

### Numerical Tests

To confirm these results, I sample from this  $d$ -dimensional Gaussian posterior (assuming  $\sigma = 1$  for simplicity) using two MH MCMC algorithms for  $n = 20,000$  iterations based on these proposal distributions. The first proposes new points assuming  $\gamma = \sqrt{2}$ . The second assumes  $\gamma = 2.5/\sqrt{d}$  in order to maintain a roughly constant acceptance fraction of 25%. As shown in [Figure 13](#), the chains behave as expected given our theoretical predictions as a function of dimensionality, with the constant proposal quickly becoming stuck while the adaptive proposal continues sampling normally. While the auto-correlation time  $\tau$  increases in both cases, the increase in the latter case (where it is driven by decreasing size/scale of the proposal distribution) is much more manageable than the former (where it is driven by the exponentially decreasing acceptance fraction).

## 8.3 MCMC with Ensemble Proposals

One drawback to the Gaussian proposals explored above is that we have to specify the structure of the distribution ahead of time. In this specific case, we assumed that:

1. the width of the posterior in each dimension (parameter) was constant such that  $\sigma_1 = \sigma_2 = \dots = \sigma_n = \sigma$  and
2. the parameters were entirely uncorrelated with each other such that the correlation coefficient  $\rho_{ij} = 0$  between any two dimensions  $i$  and  $j$ .

In general, there is no good reason to assume that either of these are true. This means we have to also estimate the entire set of  $d(d + 1)/2$  free parameters that determine the overall covariance structure of our unknown posterior distribution. Trying to adjust the covariance structure in order to improve our sampling efficiency and decrease the auto-correlation time (see §5.3 and §6.2) becomes one of the most difficult parts of running MCMC algorithms in practice.

While there are schemes to perform these adjustments during an extended burn-in period (see, e.g., [Brooks et al. 2011](#)), there is significant appeal in methods that can “auto-tune” without much additional input from the user. One class of such approaches are known as ensemble or particle methods. These methods attempt to use many  $m$  chains running simultaneously (i.e. in parallel) to improve the performance of any individual chain.

We explore three variations of ensemble methods here that attempt to exploit  $m \gtrsim d(d + 1)/2$  chains running simultaneously:

1. using the ensemble of particles to condition a Gaussian proposal distribution,
2. using trajectories from multiple particles along with Gaussian “jitter”, and
3. using affine-invariant transformations of trajectories from multiple particles.

A schematic illustration of these methods is shown in [Figure 14](#).

As we might expect, an immediate drawback of these methods is they rely on having enough particles to characterize the overall structure of the space (i.e. the curse of dimensionality). While this limits their utility when sampling from high-dimensional spaces, they can be attractive options in moderate-dimensional spaces ( $d \lesssim 25$ ) where a few hundred particles are often sufficient to ensure reasonable performance.

### 8.3.1 Gaussian Proposal

The first approach is simply a modified Gaussian proposal: at any iteration  $i$  for any chain  $j$ , we propose a new position  $\Theta_{i+1}^j$  based on the current position  $\Theta_i^j$  using a Gaussian proposal

$$\mathcal{Q}_\gamma^j(\Theta_{i+1}^j | \Theta_i^j) \propto \exp \left[ -\frac{1}{2} (\Theta_{i+1}^j - \Theta_i^j)^T (\gamma^2 \mathbf{C}_i^j)^{-1} (\Theta_{i+1}^j - \Theta_i^j) \right] \quad (89)$$

where  $T$  is the transpose operator and

$$\mathbf{C}_i^j = \text{Cov} [\{\Theta_i^1, \dots, \Theta_i^{j-1}, \Theta_i^{j+1}, \dots, \Theta_i^m\}] \quad (90)$$

is the empirical covariance matrix estimated from the current positions of the  $m$  chains excluding chain  $j$ . We repeat this process for each of the  $m$  chains in turn.

In other words, at each iteration  $i$  we want to update all  $m$  chains. We do so by updating each chain  $j$  in turn based on what the other chains are currently doing. Assuming the current position of each chain is distributed following the underlying posterior  $\mathcal{P}(\Theta)$ , it is straightforward to show that  $\mathbf{C}_i^j$  is a reasonable approximation to the unknown covariance structure of our posterior. In addition, because we exclude  $j$  when computing  $\mathbf{C}_i^j$ , this proposal is symmetric going from  $\Theta_i^j \rightarrow \Theta_{i+1}^j$  and from  $\Theta_{i+1}^j \rightarrow \Theta_i^j$ . This means that we satisfy detailed balance and do not have to incorporate any proposal-dependent factors when computing the transition probability.

### 8.3.2 Ensemble Trajectories with a Gaussian Proposal

The approach taken in §8.3.1 solves the problem of trying to tune the covariance of our initial Gaussian proposal. However, it still assumes that a Gaussian proposal is the optimal solution. A more general approach is one that does not rely on assuming a proposal explicitly, but rather only relies on the distribution of the remaining particles.

One such approach used in the literature is Differential Evolution MCMC (DE-MCMC; Storn & Price, 1997; Ter Braak, 2006). The main idea behind DE-MCMC is to rely on the relative positions of the chains at a given iteration  $i$  when making new proposals. We first randomly select two other particles  $k$  and  $l$  where  $\Theta_i^j \neq \Theta_i^k \neq \Theta_i^l$ . We then propose a new

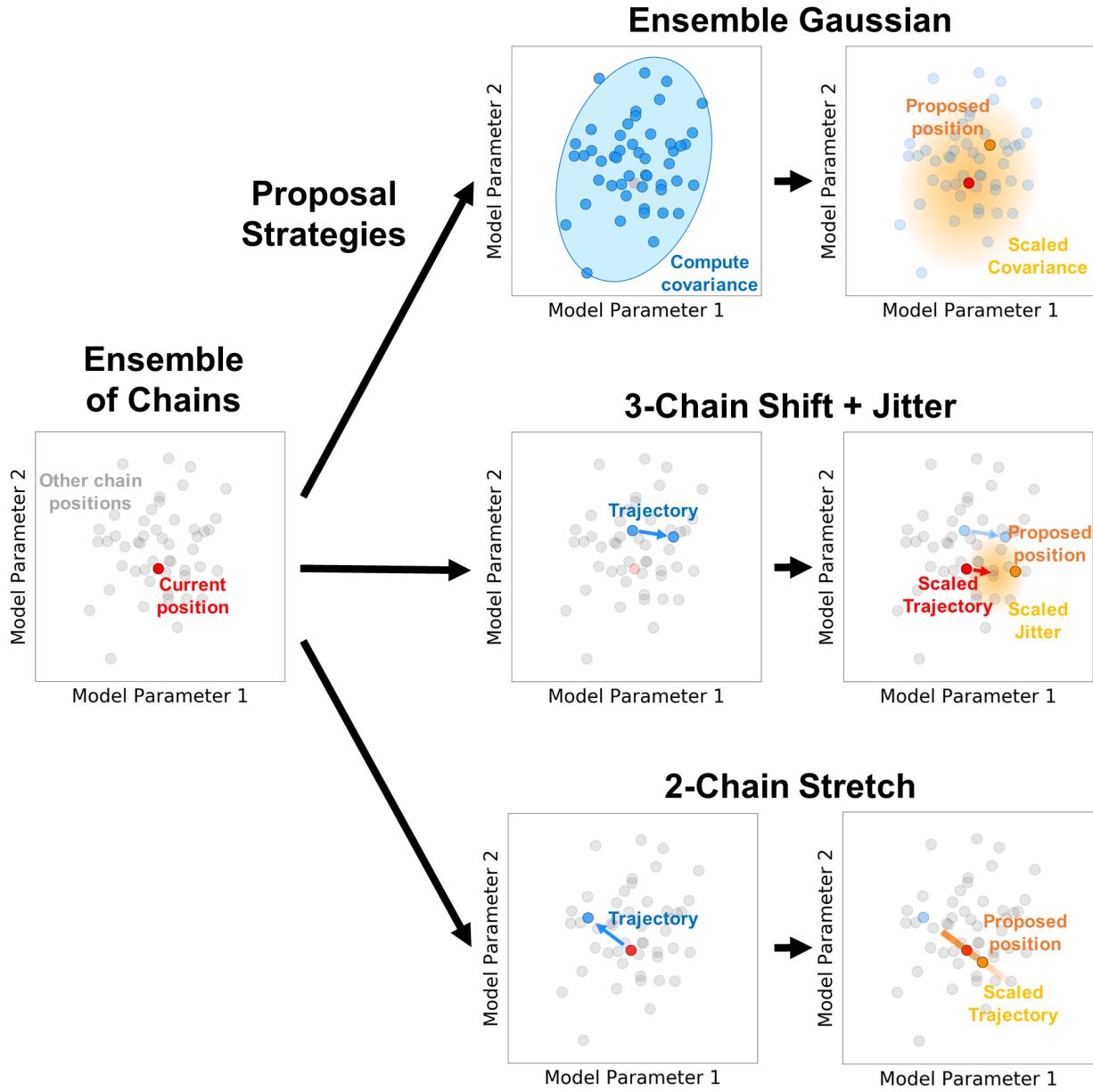


Рис. 14: A schematic illustration of the three ensemble MCMC methods described in §8.3. The current state of the chain we are interested in updating (red) and the other chains in the ensemble (gray) are shown on the left. In the top panels (ensemble Gaussian; §8.3.1), we compute the covariance of the other  $k \neq j$  chains (middle) and use a scaled version to subsequently propose a new position. In the middle panels (3-chain shift + jitter; §8.3.2), we use two additional chains  $k \neq l \neq j$  to compute a trajectory. We then propose a new position based on this scaled trajectory plus a small amount of “jitter”. In the bottom panels (2-chain stretch; §8.3.3), we use only one additional chain  $k \neq j$  to propose a new trajectory. We then propose a random position along a scaled version of this trajectory with the proposal probability varying as a function of scale. See §8.3 for additional details.

position based on the vector distance between the other two particles  $\Theta_i^k - \Theta_i^l$  with some scaling  $\gamma$  along with some additional “jitter”  $\epsilon$ :

$$\Theta_{i+1}^j = \Theta_i^j + \gamma \times (\Theta_i^k - \Theta_i^l + \epsilon) \quad (91)$$

In the case where the behavior of chains  $k$  and  $l$  are approximately independent of each other and assuming the underlying posterior distribution  $\mathcal{P}(\Theta)$  is Gaussian with some unknown mean  $\mu$  and covariance  $\mathbf{C}$  (and “standard deviation”  $\mathbf{C}^{1/2}$ ), it is straightforward to show that the distribution of  $\Theta_i^k - \Theta_i^l$  will then follow

$$\Theta_i^k - \Theta_i^l \sim \mathcal{N} [\mathbf{0}, (2\mathbf{C})^{1/2}] \quad (92)$$

Typically, the jitter  $\epsilon$  is chosen to also be Gaussian distributed with covariance  $\mathbf{C}_\epsilon$  such that

$$\epsilon \sim \mathcal{N} [\mathbf{0}, \mathbf{C}_\epsilon^{1/2}] \quad (93)$$

In general,  $\mathbf{C}_\epsilon$  is mostly used to try and avoid issues caused by finite particle sampling: since the number of unique trajectories (ignoring symmetry) is

$$n_{\text{traj}} = \binom{m-1}{2} = \frac{(m-1)!}{2!(m-3)!} = \frac{(m-1)(m-2)}{2}$$

if  $m$  is sufficiently small the DE-MCMC procedure can only explore a small number of possible trajectories at any given time, leading to extremely inefficient sampling.

Combined, this implies that the proposed position has a distribution of

$$\Theta_{i+1}^j \sim \mathcal{N} [\Theta_i^j, \gamma \times (2\mathbf{C} + \mathbf{C}_\epsilon)^{1/2}] \quad (94)$$

This shows that the 3-particle DE-MCMC procedure can generate new positions in a manner analogous to the ensemble Gaussian proposal we first discussed.

### 8.3.3 Affine-Invariant Transformations of Ensemble Trajectories

Another approach used in the literature (e.g., emcee; Foreman-Mackey et al., 2013) is the Affine-Invariant “stretch move” from Goodman & Weare (2010). This uses only one additional particle  $\Theta_i^k$  rather than two:

$$\Theta_{i+1}^j = \Theta_i^k + \gamma \times (\Theta_i^j - \Theta_i^k) \quad (95)$$

In place of the jitter term  $\epsilon$  from DE-MCMC, the stretch move instead injects some amount of randomness by allowing  $\gamma$  to vary. By sampling  $\gamma$  from some probability distribution  $g(\gamma)$ , we allow the proposals to explore various “stretches” of the direction vector. As shown in Goodman & Weare (2010), if this function is chosen such that

$$g(\gamma^{-1}) = \gamma \times g(\gamma) \quad (96)$$

then this proposal is symmetric. Typically,  $g(\gamma)$  is chosen to be

$$g(\gamma|a) = \begin{cases} \gamma^{-1/2} & a^{-1} \leq \gamma \leq a \\ 0 & \text{otherwise} \end{cases} \quad (97)$$

where  $a = 2$  is often taken as a typical value. Note that when  $\gamma = 1$ , this move leaves  $\Theta_{i+1}^j = \Theta_i^j$  unchanged.

Compared to DEMCMC, the stretch move appears to have one clear advantage: it doesn't have any reliance on some "jitter" term  $\epsilon$  that reintroduces scale-dependence into the proposal. That makes the proposal invariant to affine transformations and only sensitive to a single parameter  $a$ , which governs the range of scales the stretch factor  $\gamma$  is allowed to explore.

This lack of jitter, however, is not substantially advantageous in practice. As noted in §8.3.2,  $\epsilon$  is really designed to avoid possible degeneracies due to the limited number of available trajectories. In that case we had  $(m - 1)(m - 2)/2 \sim m^2/2$  possible trajectories; here, however, we only have  $m$  (since  $\Theta_i^j$  is always included). This is a much smaller number of possible trajectories at a given  $m$ , making this particular proposal more susceptible to that particular effect.

In addition, because this proposal involves adjusting  $\gamma$  and therefore the length of the trajectory itself, we need to consider how changing  $\gamma$  affects the total volume of the sphere centered on  $\Theta_i^j$  with radius  $\Theta_i^k - \Theta_i^j$ . As discussed in §7.2, the differential volume increases as  $r^{d-1}$ . Therefore, increasing or decreasing  $\gamma$  substantially adjusts the differential volume in our proposal. This involves introducing a steep boost/penalty into our transition probability, which now becomes:

$$T(\Theta_{i+1}^j | \Theta_i^j, \gamma) = \min \left[ 1, \gamma^{d-1} \frac{\mathcal{P}(\Theta_{i+1}^j)}{\mathcal{P}(\Theta_i^j)} \right] \quad (98)$$

This heavily favors proposals with  $\gamma > 1$  (outwards) and heavily disfavors proposals with  $\gamma < 1$  as  $d$  increases to account for the exponentially increasing volume at larger radii.

Finally, while this stretch move actually generates proposals in the right overall direction, it is not efficient at generating samples within the bulk of the posterior mass as the dimensionality increases. As discussed in §8.2, given the typical position of  $\Theta_i^j$ , the typical length-scale of the proposed positions needs to shrink by  $\propto 1/\sqrt{d}$  in order to guarantee our new sample remains within the bulk of the posterior mass. However, the form for  $g(\gamma|a)$  specified above instead ensures that  $\gamma$  will always be between  $1/a$  and  $a$ . Even if we attempt to account for this effect by letting  $a(d) \rightarrow 1$  as  $d \rightarrow \infty$  in order to target a constant acceptance fraction and ensure more overlap, the asymmetry of our proposal and the  $\gamma^{d-1}$  term in the transition probability systematically biases our proposed and accepted positions compared with the ideal distribution. This subsequently leads to larger auto-correlation times, mostly counteracting any expected gains.

## Numerical Tests

To confirm these results, I sample from this  $d$ -dimensional Gaussian posterior (assuming  $\sigma = 1$  for simplicity) using each of these ensemble MH MCMC algorithms with  $n = 1500$  iterations with  $m = 100$  chains. In the first case, I propose a new position for chain  $j$  at iteration  $i$  using a Gaussian distribution with a covariance  $\gamma^2 \mathbf{C}_i^j$  computed over the remaining ensemble of  $k \neq j$  chains, where the scale factor  $\gamma = 2.5/\sqrt{d}$  is chosen to target a constant acceptance fraction of roughly 25%. In the second case, I propose new positions using the DE-MCMC algorithm with a scale factor of  $\gamma = 1.7/\sqrt{d}$  and additional Gaussian jitter with

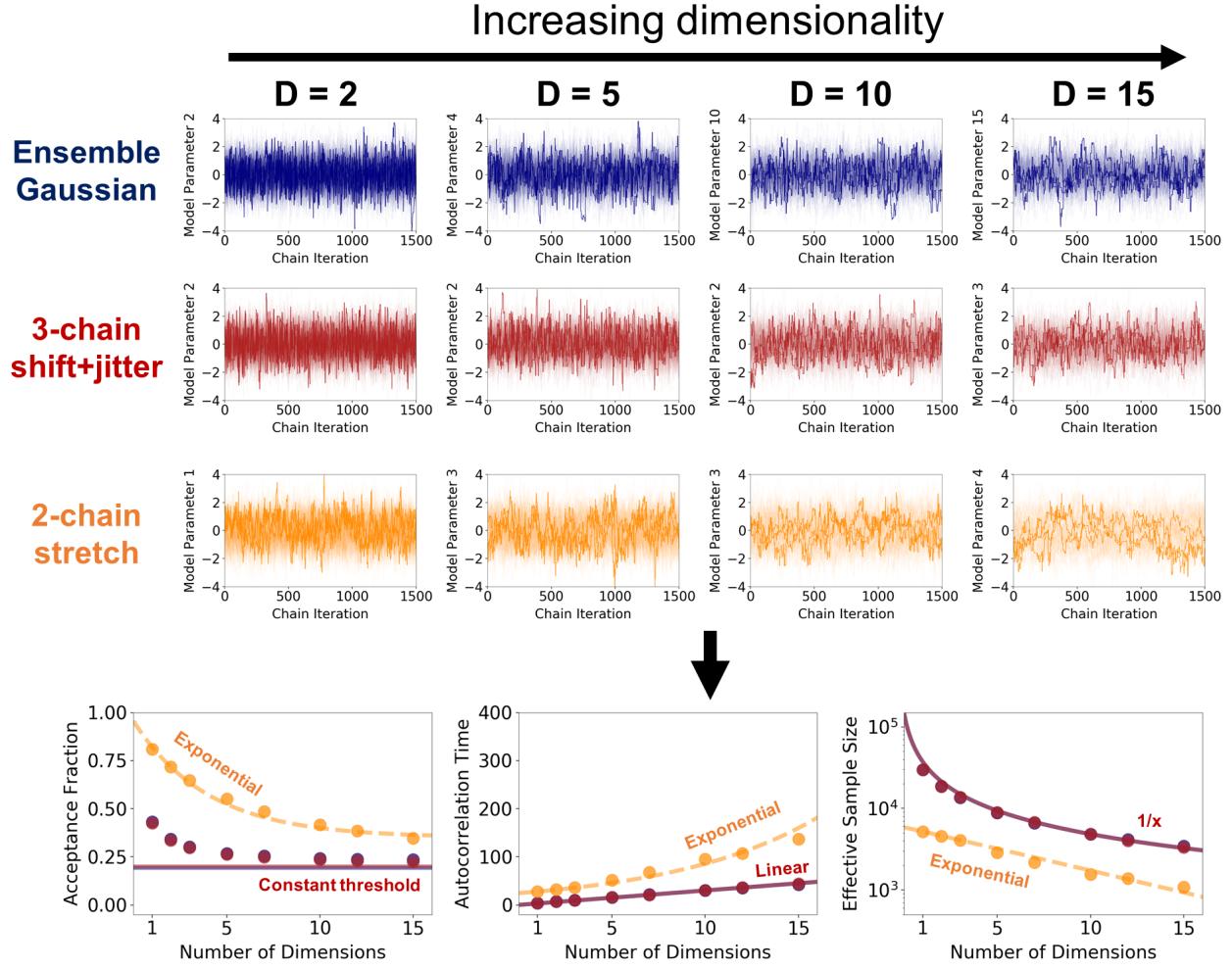


Рис. 15: Numerical results showcasing the performance of several ensemble MH MCMC samplers on our toy problem, a  $d$ -dimensional Gaussian with mean  $\mu = 0$  and standard deviation  $\sigma = 1$  in every dimension. The top series of panels show snapshots of a random parameter from the collection of chains (with a few chains highlighted) as a function of dimensionality (increasing from left to right) assuming ensemble Gaussian proposals with  $\gamma = 2.5/\sqrt{d}$  (blue), 3-chain “shift and jitter” proposals with  $\gamma = 1.7/\sqrt{d}$  (red), and 2-chain “stretch” proposals with  $\gamma$  drawn from the distribution  $g(\gamma|a)$  with  $a = 2$  as described in §8.3.3 (orange). The bottom panels show the corresponding acceptance fractions (left), auto-correlation times (middle), and effective sample sizes (right) from our chains (colored points) as a function of dimensionality. Approximations based on the §8.2 are shown as light solid colored lines, with dashed lines showing rough fits. The first two methods, which allow the size of the proposal to shrink, are able to propose samples within the bulk of the posterior mass. The last method, which is unable to do so, instead proposes exponentially fewer good positions as the dimensionality increases. See §8.3 for additional details.

covariance  $\mathbf{C}_\epsilon = \mathbf{C}_i^j / 5$  derived from the remaining chains in the ensemble, again targeting an acceptance fraction of roughly 25%. In the third case, I propose new positions using the affine-invariant stretch move assuming the typical form for  $g(\gamma|a)$  with  $a = 2$ .<sup>7</sup>

As shown in [Figure 15](#), the chains behave as expected given our theoretical predictions as a function of dimensionality. Similar to the adaptive Gaussian case, the first two approaches continue sampling efficiently even as  $d$  increases. The affine-invariant stretch move, however, experiences exponentially-decreasing efficiency and struggles to sample the posterior effectively.

## 8.4 Additional Comments

Before concluding, I wish to emphasize that the toy problem explored in this section should only be interpreted as a tool to build intuition surrounding how certain methods are expected to behave in a controlled environment. While the behavior as a function of dimensionality helps to illustrate common issues, in practice the performance of any method will depend on the specific problem, tuning parameters, the time spent on tuning, and many other possible factors. Since it is always possible to find problems for which any particular method will perform well or poorly, I encourage users to try out a variety of approaches to find the ones that work best for their problems.

## 9 Conclusion

Bayesian statistical methods have become increasingly prevalent in modern scientific analysis as models have become more complex. Exploring the inferences we can draw from these models often requires the use of numerical techniques, the most popular of which is known as Markov Chain Monte Carlo (MCMC).

In this article, I provide a conceptual introduction to MCMC that seeks to highlight the what, why, and how of the overall approach. I first give an overview of Bayesian inference and discuss what types of problems Bayesian inference generally is trying to solve, showing that most quantities we are interested in computing require integrating over the posterior density. I then outline approaches to computing these integrals using grid-based approaches, and illustrate how adaptively changing the resolution of the grid naturally transitions into the use of Monte Carlo methods. I illustrate how different sampling strategies affect the overall efficiency in order to motivate why we use MCMC methods. I then discuss various details related to how MCMC methods work and examine their expected overall behavior based on simple arguments derived from how volume and posterior density behave as the number of parameters increases. Finally, I highlight the impact this conceptual understanding has in practice by comparing the performance of various MCMC methods on a simple toy problem.

I hope that the material in this article, along with the exercises and applications, serve as a useful resource that helps build up intuition for how MCMC and other Monte Carlo methods work. This intuition should be helpful when making decisions over when to apply MCMC methods to your own problems over possible alternatives, developing novel proposals

---

<sup>7</sup>Allowing  $a(d)$  to vary as a function of dimensionality to target a roughly constant acceptance fraction gives similar results.

and sampling strategies, and characterizing what issues you might expect to encounter when doing so.

## Acknowledgements

JSS is grateful to Rebecca Bleich for continuing to tolerate his (over-)enthusiasm for sampling during their time together. He would also like to thank a number of people for helping to provide much-needed feedback during earlier stages of this work, including Catherine Zucker, Dom Pesce, Greg Green, Kaisey Mandel, Joel Leja, David Hogg, Theron Carmichael, and Jane Huang. He would also like to thank Ana Bonaca, Charlie Conroy, Ben Cook, Daniel Eisenstein, Doug Finkbeiner, Boryana Hadzhiyska, Will Handley, Locke Patton, and Ioana Zelko for helpful conversations surrounding the material.

JSS also wishes to thank Kaisey Mandel and the Institute of Astronomy at the University of Cambridge, Hans-Walter Rix and the Galaxies and Cosmology Department at the Max Planck Institute for Astronomy, and Renée Hložek, Bryan Gaensler, and the Dunlap Institute for Astronomy and Astrophysics at the University of Toronto for their kindness and hospitality while hosting him over the period where a portion of this work was being completed.

JSS acknowledges financial support from the National Science Foundation Graduate Research Fellowship Program (Grant No. 1650114) and the Harvard Data Science Initiative.

## Список литературы

Asmussen, S., & Glynn, P. W. 2011, Statistics & Probability Letters, 81, 1482 , doi: <https://doi.org/10.1016/j.spl.2011.05.004>

Blitzstein, J., & Hwang, J. 2014, Introduction to Probability, Chapman & Hall/CRC Texts in Statistical Science (CRC Press/Taylor & Francis Group). <https://books.google.com/books?id=ZwSlMAEACAAJ>

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. 2011, Handbook of Markov Chain Monte Carlo (CRC press)

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Pub. of the Astron. Soc. of the Pac., 125, 306, doi: 10.1086/670067

Gelman, A., Carlin, J., Stern, H., et al. 2013, Bayesian Data Analysis, Third Edition, Chapman & Hall/CRC Texts in Statistical Science (Taylor & Francis). <https://books.google.com/books?id=ZXL6AQAAQBAJ>

Gelman, A., & Meng, X.-L. 1998, Statist. Sci., 13, 163, doi: 10.1214/ss/1028905934

Gelman, A., & Rubin, D. B. 1992, Statistical Science, 7, 457, doi: 10.1214/ss/1177011136

Goodman, J., & Weare, J. 2010, Communications in Applied Mathematics and Computer Science, 5, 65, doi: 10.2140/camcos.2010.5.65

- Hastings, W. 1970, Biometrika, 57, 97, doi: 10.1093/biomet/57.1.97
- Hogg, D. W., & Foreman-Mackey, D. 2018, The Astrophys. Journal Supp., 236, 11, doi: 10.3847/1538-4365/aab76e
- Kish, L. 1965, Survey sampling, Wiley classics library (J. Wiley). <https://books.google.com/books?id=xiZmAAAAIAAJ>
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, Journal of Chem. Phys., 21, 1087, doi: 10.1063/1.1699114
- Neal, R. M. 2012, arXiv e-prints, arXiv:1206.1901. <https://arxiv.org/abs/1206.1901>
- Skilling, J. 2004, in American Institute of Physics Conference Series, Vol. 735, American Institute of Physics Conference Series, ed. R. Fischer, R. Preuss, & U. V. Toussaint, 395–405
- Skilling, J. 2006, Bayesian Anal., 1, 833, doi: 10.1214/06-BA127
- Storn, R., & Price, K. 1997, Journal of global optimization, 11, 341
- Ter Braak, C. J. 2006, Statistics and Computing, 16, 239
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. 2019, arXiv e-prints, arXiv:1903.08008. <https://arxiv.org/abs/1903.08008>