Abstract

The Gaussian distribution has many useful properties. Yet there are few resources that derive these properties from scratch in a concise and comprehensive manner. This technical note is an ongoing effort to develop such a resource. The statements are written as generally as possible, with clean and accessible proofs whenever applicable. Some novel extensions of existing results are provided (e.g., to multivariate forms, non-iid noises). Equipped with the results in this note, we are able to derive complex methods such as diffusion models and sparse Gaussian processes with relative ease, for instance by simply invoking the Gaussian chain rule and Bayes’ rule instead of calculating any integral.

Contents

1 Definitions 3

2 Basic Properties 3
  2.1 Shape . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
  2.2 Linear Transformation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
  2.3 Log-Likelihood . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
  2.4 Sample Mean and Covariance . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

3 Joint Distribution 4
  3.1 Linear Combinations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
  3.2 Chain Rule . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
  3.3 Bayes’ Rule . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

4 Entropy 5
  4.1 Mutual Information . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
    4.1.1 Additive white Gaussian noise channel . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
    4.1.2 Correlated standard normal channel . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

5 Central Limit Theorem 6

6 Exponential Family 7
  6.1 Exponential Tilting . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
    6.1.1 Aside: Tweedie’s formula . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

7 Sub-Gaussian Distributions 8

8 Cumulative Distribution Function 8

9 Gaussian Processes 9
  9.1 Sparse GPs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

10 TODO: High-Dimensional Behavior 11
A Integration 12
A.1 Single-Variable .......................................................... 12
  A.1.1 Substitution in practice ............................................. 12
A.2 Multi-Variable .......................................................... 13
  A.2.1 Applications to probability ...................................... 13

B Continuous Entropy and KL Divergence 14
B.1 Shortcomings of Continuous Entropy .............................. 15
  B.1.1 Inconsistency with Shannon entropy .......................... 15
  B.1.2 Variability under change of coordinates ...................... 16

C Moment-Generating Function 17
C.1 Cumulant-Generating Function .................................... 18

D Exponential Family 19
  D.1 Exponential Tilting .................................................. 19
  D.2 Unnormalized Form .................................................. 19
    D.2.1 Discussions ..................................................... 20
  D.3 Tweedie’s Formula ................................................ 21

E Laplace Approximation 22

F Linear Regression 22
  F.1 Bayesian Linear Regression ...................................... 22

G Lemmas 23

H Individually Normal But Not Jointly Normal 39
1 Definitions

Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}_{\geq 0}$. We assert $\Sigma > 0$ (i.e., symmetric and positive-definite) to avoid handling degenerate cases. The **Gaussian distribution** is a mapping $\mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to [0, 1]$ defined as

$$\mathcal{N}(\mu, \Sigma)(x) := \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right) \tag{1}$$

which integrates to 1 over $\mathbb{R}^d$ (Lemma G.4) and is thus a valid probability distribution. The distribution has the moment-generating function (MGF) of $M_X(t) = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t)$ (Appendix C), which readily shows that $\mu$ is the mean and $\Sigma$ the covariance. If $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, (1) is a product distribution of univariate $\mathcal{N}(\mu_i, \sigma_i^2)$. The **standard Gaussian** is the special case with $\mu = 0_d$ and $\Sigma = I_d$. The following statements about a random variable $X \in \mathbb{R}^d$ are equivalent (Lemma G.16):

1. $X \sim \mathcal{N}(\mu, \Sigma)$. That is, the probability of $X = x$ is $\mathcal{N}(\mu, \Sigma)(x)$ defined in (1).
2. The MGF of $X$ is $M_X(t) = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t)$.
3. $X = \mu + \Sigma^{1/2} Z$ where $Z \sim \mathcal{N}(0_d, I_d)$.
4. $a^\top X \sim \mathcal{N}(a^\top \mu, a^\top \Sigma a)$ for all nonzero $a \in \mathbb{R}^d$.
5. The log probability of $X = x$ is equal to $-\frac{1}{2} x^\top \Sigma^{-1} x + (\Sigma^{-1} \mu)^\top x + C$ where $C \in \mathbb{R}$ is constant in $x$.

If any holds, we say $X \in \mathbb{R}^d$ is **normally distributed** with parameters $(\mu, \Sigma)$. Note that 3 and 4 just reduce general normality to simpler forms of (1) (standard and univariate). These alternative definitions are useful in different contexts, for instance

- 2 shows that a point-mass distribution on $x \in \mathbb{R}^d$ is “normal” with parameters $(x, 0_{d \times d})$, since its MGF is $\mathbb{E}[\exp(t^\top X)] = \exp(t^\top x)$.
- 3 is the popular Gaussian reparameterization trick where we view $X$ as a perturbation of $(\mu, \Sigma)$.
- 4 is handy when showing that $Y$ and $Z$ are jointly normal (Section 3): it is sufficient to show that any scalar projection of $(Y, Z)$ using a nonzero vector is (univariate) normal.
- 5 “completes the square” for you. By putting the log probability of $X$ in this form, we show that $X$ is normal and identify its covariance and mean by matching the second- and first-order terms.

2 Basic Properties

2.1 Shape

(1) implies that the distribution is symmetric: $\mathcal{N}(\mu, \Sigma)(x) = \mathcal{N}(\mu, \Sigma)(-x)$. The gradient and the Hessian matrix of $\mathcal{N}(\mu, \Sigma)$ at $x \in \mathbb{R}^d$ are

$$\left( \nabla \mathcal{N}(\mu, \Sigma) \right)(x) = -\mathcal{N}(\mu, \Sigma)(x) \times \Sigma^{-1}(x - \mu)$$

$$\left( \nabla^2 \mathcal{N}(\mu, \Sigma) \right)(x) = -\mathcal{N}(\mu, \Sigma)(x) \times (\Sigma^{-1} - \Sigma^{-1} (x - \mu)(x - \mu)^\top \Sigma^{-1})$$

where the Hessian is negative-definite at $x = \mu$ but can be indefinite at other points (Lemma G.32). The distribution is not concave, but it is (strictly) log-concave, thus quasiconcave, with $\mu$ as the unique mode (as well as the mean).

2.2 Linear Transformation

A critical property of the Gaussian distribution is that it is closed under linear transformation. Note that definitions 3 and 4 are consistent with this property. For any $A \in \mathbb{R}^{d' \times d}$ and $b \in \mathbb{R}^{d'}$ where $A$ is full-rank with $d' \leq d$ (so that $A \Sigma A^\top > 0$), $X \sim \mathcal{N}(\mu, \Sigma)$ implies (Lemma C.2):

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top) \tag{2}$$
2.3 Log-Likelihood

The Gaussian log-likelihood with a Gaussian random mean is again a Gaussian log-likelihood with a regularization. Pick any $A \in \mathbb{R}^{d \times d}$ and $\Omega \in \mathbb{R}_{>0}^{d \times d}$. For all $y \in \mathbb{R}^d$ (Lemma G.10),

$$\mathbb{E}_{X \sim \mathcal{N}(\mu, \Sigma)} \left[ \log \mathcal{N}(AX, \Omega)(y) \right] = \log \mathcal{N}(A\mu, \Omega)(y) - \frac{1}{2} \text{tr} (\Omega^{-1}A\Sigma A^\top)$$ (3)

2.4 Sample Mean and Covariance

Another characteristic of the Gaussian distribution is that the sample mean and covariance are independent. For any iid $X_1 \ldots X_N \sim \text{Unk}$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}_{>0}^{d \times d}$, unbiased estimators of the mean and covariance are given by

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i \quad \text{and} \quad \bar{S}_N^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X}_N)(X_i - \bar{X}_N)^\top$$

It turns out that $\bar{X}_N$ and $\bar{S}_N^2$ are independent iff $\text{Unk}$ is normal (Geary, 1936). In fact, if $\text{Unk}$ is normal, then $\bar{X}_N \sim \mathcal{N}(\mu, (1/N)\Sigma)$ and, independently, $(N-1)\bar{S}_N^2 \sim \mathcal{W}_d(N-1, \Sigma)$ where $\mathcal{W}_d$ is known as the Wishart distribution (proof). If $d = 1$ and $\Sigma = \sigma^2 I$, this implies the better known form $(N-1)/\sigma^2 \bar{S}_N^2 \sim \chi^2(N-1)$ where $\chi^2(k)$ is the chi-square distribution with $k$ degrees of freedom.

3 Joint Distribution

We say $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are jointly normally distributed with parameters $(\mu, \Sigma)$ if the concatenation $(X, Y)$ follows $\mathcal{N}(\mu, \Sigma)$. More explicitly,

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \right)$$

where $\mu_X \in \mathbb{R}^d$, $\mu_Y \in \mathbb{R}^{d'}$, $\Sigma_X \in \mathbb{R}_{>0}^{d \times d}$, $\Sigma_Y \in \mathbb{R}_{>0}^{d' \times d'}$, $\Sigma_{XY} \in \mathbb{R}^{d \times d'}$, and $\Sigma_{YX} = \Sigma_{XY}^\top$. A subtle fact is that $X$ and $Y$ can be individual normal but not jointly normal (Appendix H), so we must explicitly establish joint normality even for normal variables (e.g., by using 4 or 5). If $X$ and $Y$ are individually normal and independent, then they are jointly normal since we can write

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{bmatrix} \right)$$

If $X, Y$ are jointly normal, uncorrelatedness implies independence (thus they are equivalent). But we must show joint normality before claiming independence from uncorrelatedness. For instance, Appendix H gives $X, Y \in \mathbb{R}$ that are individually normal (but not jointly normal) and uncorrelated, but not independent. The following results are useful when inferring independence from uncorrelatedness:

$$\forall A \in \mathbb{R}^{n \times d}, B \in \mathbb{R}^{m \times d} : \quad A\Sigma B^\top = 0_{n \times m} \iff AX \in \mathbb{R}^n \text{ and } BX \in \mathbb{R}^m \text{ are independent} \quad (4)$$

$$\forall A, B \in \mathbb{R}^{d \times d} : \quad A\Sigma B = 0_{d \times d} \iff X^\top AX \in \mathbb{R} \text{ and } X^\top BX \in \mathbb{R} \text{ are independent} \quad (5)$$

where $X \sim \mathcal{N}(\mu, \Sigma)$. Despite their striking similarity, the linear form (4) is simple to prove (Lemma G.11) but the quadratic form (5), known as Craig’s theorem (Craig, 1943), is surprisingly difficult and has a long and complicated history (Driscoll and Gundberg Jr., 1986).

---

1 Specifically, $\mathcal{W}_d(k, \Sigma)$ is the distribution over $(u_1 \ldots u_k)^\top (u_1 \ldots u_k) \in \mathbb{R}^{d \times d}$ where $u_1 \ldots u_k \in \mathbb{R}^d$ are iid samples from $\mathcal{N}(0_d, \Sigma)$.

2 We must have $\Sigma_X, \Sigma_Y \succ 0$ since they are main-diagonal blocks of $\Sigma \succ 0$ (Lemma G.9) and $\Sigma_{XY} = \Sigma_{YX}^\top$ since $\Sigma$ is symmetric.

3 This follows from the form of the conditional distribution (7):

$$\Sigma_{XY} = 0_{d \times d'} \implies \mathcal{N}(\mu, \Sigma)(y|x) = \mathcal{N}(\mu_Y + \Sigma_{YX} \Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY})(y) = \mathcal{N}(\mu_Y, \Sigma_Y)(y)$$
3.1 Linear Combinations

Let $A \in \mathbb{R}^{p \times d}$, $B \in \mathbb{R}^{p \times d'}$, and $b \in \mathbb{R}^p$ where $A, B$ are full-rank with $p \leq \min(d, d')$. If $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are jointly normal with parameters $(\mu, \Sigma)$, we have from (2) that

$$AX + BY + b \sim \mathcal{N}(A\mu_X + B\mu_Y + b, A\Sigma_X A^\top + A \Sigma_{XY} B^\top + B \Sigma_{Y \cdot} A^\top + B \Sigma_Y B^\top)$$

(6)

In particular, if $X$ and $Y$ are independently normal, then their sum is distributed as

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$$

Note that we need joint normality to guarantee the normality of a linear combination. In general a linear combination of normal variables may not be normal (e.g., (99)).

3.2 Chain Rule

If $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are jointly normal with parameters $(\mu, \Sigma)$, and if $\Sigma_Y - \Sigma_{Y \cdot} \Sigma_X^{-1} \Sigma_{XY}$ is invertible, then $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and (Lemma G.12)

$$Y | X = x \sim \mathcal{N}(\mu_Y + \Sigma_{Y \cdot X} \Sigma_X^{-1} (x - \mu_X), \Sigma_Y - \Sigma_{Y \cdot X} \Sigma_X^{-1} \Sigma_{XY})$$

(7)

(7) can be expressed more simply in terms of the blocks of the precision matrix $\Lambda = \Sigma^{-1}$. In this case (Lemma G.13),

$$Y | X = x \sim \mathcal{N}(\mu_Y - \Lambda_{Y \cdot X} (x - \mu_X), \Lambda_Y^{-1})$$

(8)

3.3 Bayes’ Rule

If $X \sim \mathcal{N}(\mu, \Sigma_X)$ (“Gaussian prior”) and $Y | X = x \sim \mathcal{N}(Ax + b, \Sigma_Y)$ (“linear-Gaussian likelihood”) where $A \in \mathbb{R}^{d' \times d}$ and $b \in \mathbb{R}^{d'}$, then

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_X A^\top \\ A \Sigma_X & A \Sigma_X + A \Sigma_{Y \cdot} A^\top \end{bmatrix}\right)$$

along with the marginal and posterior distributions

$$Y \sim \mathcal{N}(A\mu + b, \Sigma_Y + A \Sigma_X A^\top)$$

(9)

$$X | Y = y \sim \mathcal{N}(\Lambda_X^{-1} (\Sigma_X^{-1} \mu + A^\top \Sigma_Y^{-1} (y - b)), \Lambda_X^{-1})$$

(10)

where $\Lambda_X = \Sigma_X^{-1} + A^\top \Sigma_Y^{-1} A$ (Lemma G.14). In particular, the Gaussian prior is conjugate for the linear-Gaussian likelihood.

4 Entropy

Let $\mu' \in \mathbb{R}^d$ and $\Sigma' \in \mathbb{R}^{d \times d}$ be parameters of an additional Gaussian distribution over $\mathbb{R}^d$. Then the cross entropy between $\mathcal{N}(\mu', \Sigma')$ and $\mathcal{N}(\mu, \Sigma)$ is (Lemma G.6):

$$H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2} (\mu' - \mu)^\top \Sigma^{-1} (\mu' - \mu) + \frac{1}{2} \text{tr} \left( \Sigma^{-1} \Sigma' \right) + \frac{1}{2} \log((2\pi)^d \det(\Sigma))$$

(11)

This is sufficient to derive entropy and KL divergence:

$$H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \log\left((2\pi e)^d \det(\Sigma)\right)$$

(12)

$$\text{KL}(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2} (\mu' - \mu)^\top \Sigma^{-1} (\mu' - \mu) + \frac{1}{2} \text{tr} \left( \Sigma^{-1} \Sigma' - I_{d \times d} \right) + \frac{1}{2} \log\left(\frac{\det(\Sigma)}{\det(\Sigma')}\right)$$

(13)

Notably, $\mathcal{N}(\mu, \Sigma)$ has the largest entropy among all distributions over $\mathbb{R}^d$ with mean $\mu$ and covariance $\Sigma$ (Theorem B.1). This is mainly because it standardizes $x$ inside the exponential function.
4.1 Mutual Information

Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$ be jointly normal with parameters $(\mu, \Sigma)$. If $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible, then conditional entropy and mutual information are (Lemma G.15):

$$H(Y|X = x) = \frac{1}{2}\log \left( 2\pi e \right) \det(\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})$$

$$I(X,Y) = \frac{1}{2}\log \left( \frac{\det(\Sigma_X)\det(\Sigma_Y)}{\det(\Sigma)} \right)$$

where $x \in \mathbb{R}^d$ is arbitrary (so $H(Y|X) = H(Y|X = x)$). Note that $I(X,Y)$ is infinite if $Y = X$. By the noisy-channel coding theorem, mutual information is the capacity (highest information rate that can be achieved nearly error-free) of a communication channel between $X$ and $Y$. Below we give some well-known models with controllable mutual information.

4.1.1 Additive white Gaussian noise channel

Let $X \sim \mathcal{N}(0, \sigma^2)$ and $Z \sim \mathcal{N}(0, \nu^2)$ independently, and define $Y = X + Z$. $X$ and $Y$ are jointly normal because $a_1X + a_2Y = (a_1 + a_2)X + a_2Z$ is a sum of independently normal variables and thus normal for all nonzero $a = (a_1, a_2)$ (definition 4). Since $\text{Var}(Y) = \sigma^2 + \nu^2$ and $\text{Cov}(X,Y) = \sigma^2$,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \nu^2 \end{pmatrix} \right) \implies I(X,Y) = \frac{1}{2}\log \left( 1 + \frac{\sigma^2}{\nu^2} \right)$$

Thus $I(X,X+Z)$ grows logarithmically in signal-to-noise ratio $\frac{\sigma^2}{\nu^2}$.

4.1.2 Correlated standard normal channel

Let $X, Y \in \mathbb{R}$ be jointly standard normal with correlation $\rho < 1$. One way to construct them is to let $X, Z \overset{iid}{\sim} \mathcal{N}(0,1)$ and set $Y = \rho X + \sqrt{1-\rho^2}Z$. Then

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \implies I(X,Y) = -\frac{1}{2}\log (1 - \rho^2)$$

By taking the correlation $\rho \to 1$ we can arbitrarily increase $I(X,Y)$.

5 Central Limit Theorem

Let $\text{Unk}(\mu, \sigma^2)$ denote an unknown distribution over $\mathbb{R}$ with mean $\mu$ and variance $\sigma^2 > 0$. It is often of interest to consider the sample average $\bar{X}_N$ defined as

$$X_1 \ldots X_N \overset{iid}{\sim} \text{Unk}(\mu, \sigma^2) \quad \bar{X}_N := \frac{1}{N} \sum_{i=1}^{N} X_i$$

The average is itself random: every time we draw $N$ iid samples from Unk($\mu, \sigma^2$), we draw a single sample of $\bar{X}_N$. We can easily verify that $\mathbb{E}[\bar{X}_N] = \mu$ and $\text{Var}(\bar{X}_N) = \frac{\sigma^2}{N}$, which states that $\bar{X}_N$ concentrates around $\mu$ as $N \to \infty$ (this is called the “law of large numbers”). But what is the distribution of $\bar{X}_N$? The central limit theorem (CLT) states that $\bar{X}_N$ is asymptotically normal. More precisely, as $N \to \infty$ we have

$$\sqrt{N} (\bar{X}_N - \mu) \overset{\text{apprx.}}{\sim} \mathcal{N}(0, \sigma^2)$$

or, using the closure under linear transformation,

$$\bar{X}_N \overset{\text{apprx.}}{\sim} \mathcal{N} \left( \mu, \frac{\sigma^2}{N} \right)$$

which is consistent with but not implied by the law of large numbers. CLT allows us to make probabilistic statements about sample averages regardless of the underlying distribution. For instance, if $X_1 \ldots X_N$ are arbitrary iid samples
with mean 42 and variance 7, then approximately \( \bar{X}_N \sim \mathcal{N}(42, \frac{7}{N}) \) so that we can calculate quantities like \( \Pr(\bar{X}_N \leq 50) \) (e.g., by consulting a standard normal table).

A proof of CLT shows that the KL divergence between the distribution of \( \sqrt{N} (\bar{X}_N - \mu) \) and \( \mathcal{N}(0, \sigma^2) \) goes to zero as \( N \to \infty \). It is nontrivial: we refer to Marsh (2013) for details. CLT generalizes naturally to multivariate. If \( \text{Unk}(\mu, \Sigma) \) is an unknown distribution over \( \mathbb{R}^d \) with mean \( \mu \) and covariance \( \Sigma > 0 \), then the average \( \bar{X}_N \) of samples \( X_1 \ldots X_N \sim \text{Unk}(\mu, \Sigma) \) satisfies as \( N \to \infty \):

\[
\sqrt{N} (\bar{X}_N - \mu) \approx \mathcal{N}(0, \Sigma) \quad (18)
\]

\[
\bar{X}_N \approx \mathcal{N}(\mu_0 \frac{1}{N} \Sigma) \quad (19)
\]

## 6 Exponential Family

An exponential family is any set of distributions over \( \mathbb{R}^d \) that can be expressed as

\[
q_{h,\tau,\theta}(x) = \frac{h(x)}{\text{base measure (} > 0)} \exp \left( \frac{\theta^\top}{\text{natural parameter (} \mathbb{R}^m \text{)} \text{ sufficient statistic (} \mathbb{R}^m \text{)}} \right) \tau(x) - A_{h,\tau}(\theta) \quad (20)
\]

where the log-partition function has the important property of generating cumulants of the sufficient statistic when differentiated (e.g., \( \nabla A_{h,\tau}(\theta) \) is the mean of \( \tau(x) \) where \( x \sim q_{h,\tau,\theta} \)). The set of Gaussian distributions is an exponential family (Appendix D), with one parameterization (Lemma G.23)

\[
\mathcal{N}(\mu, \Sigma)(x) = \frac{1}{(2\pi)^d} \exp \left( \frac{-\frac{1}{2} \text{vec}^{-1}(\Sigma) \mu^\top}{\text{base measure (\mathbb{R}^{d+1}) \text{ natural parameter (} \mathbb{R}^{d(d+1)} \text{) sufficient statistic (} \mathbb{R}^{d(d+1)} \text{) \text{ log-partition function)}}} \right) \quad (21)
\]

where \( \text{vec}(A) \in \mathbb{R}^{d^2} \) vectorizes matrix \( A \in \mathbb{R}^{d \times d} \). Thus it inherits the usual properties of an exponential family such as the concavity of the likelihood function and the availability of conjugate priors.

### 6.1 Exponential Tilting

Any “base” distribution \( p(x) \) can be used as the (normalized) base measure in (20) and, using the identity sufficient statistic \( \tau(x) = x \), generates a new exponential family as \( g_t(x) \propto e^{\tau(x)} p(x) \) indexed by natural parameter \( t \in \mathbb{R}^d \). This technique is called exponential tilting. A useful fact is that the Gaussian distributions are closed under exponential tilting (Lemma G.26):

\[
\Pr(X_t = x) \propto e^{\tau(x) \times \mathcal{N}(\mu, \Sigma)(x)} \quad \Rightarrow \quad X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma) \quad (22)
\]

#### 6.1.1 Aside: Tweedie’s formula

(22) can be used to derive a score-based Bayesian estimator called Tweedie’s formula, which is a point estimator implied from the following posterior for \( \mu \sim g \) and \( x_t \sim \mathcal{N}(\mu, \Sigma) \) (Lemma D.2)

\[
\mu | x \sim \text{Unk}(x + \Sigma \nabla l(x), \Sigma (I_{d \times d} + \nabla^2 l(x) \Sigma)) \quad \text{Tweedie’s formula}
\]

where \( l(x) = \log m(x) \) is the log-marginal. It is typically motivated as correcting for “selection bias” (Efron, 2011). Suppose we observe \( N \) samples \( x_i \sim \mathcal{N}(\mu_i, \sigma^2) \) where the mean itself is drawn from some unknown prior \( \mu_i \sim g \) every time. Consider the problem of estimating the mean of \( x_{\max} = \max_{i=1}^N x_i \). Maximum-likelihood estimation \( \hat{\mu}_{\text{MLE}} = x_{\max} \) almost certainly overestimates the true mean for large \( N \). Instead, we can consider the bias-corrected estimator \( \hat{\mu} = x_{\max} + \sigma^2 \frac{\partial}{\partial \mu} (\log m(x)) \big|_{x=x_{\max}} \) where \( m(x) = \int_{\mu \in \mathbb{R}} g(\mu) \mathcal{N}(\mu, \sigma^2)(x) d\mu \) is the marginal distribution. Intuitively, we will have \( \frac{\partial}{\partial \mu} (\log m(x)) \big|_{x=x_{\max}} < 0 \) because \( x_{\max} \) is too large given the knowledge of a shared prior.
7 Sub-Gaussian Distributions

A random scalar $S \in \mathbb{R}$ with $E[S] = 0$ is **sub-Gaussian with variance factor** $\sigma^2$, denoted by $S \sim \mathcal{G}(\sigma^2)$, if

$$\psi_S(t) \leq \psi_{Z \sim \mathcal{N}(0, \sigma^2)}(t) = \frac{\sigma^2 t^2}{2}$$

(23)

for all $t \in \mathbb{R}$. It is stable in the following sense:

1. $\text{Var}(S) \leq \sigma^2$ (Lemma G.29).
2. $-S \sim \mathcal{G}(\sigma^2)$. This can be seen by noting that $\psi_{-S}(t) = \psi_S(-t)$.
3. $\Pr(S \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$ for all $\epsilon \geq 0$. Use Chernoff’s inequality (G.19) with Lemma G.30 and (70).
4. If $S_1 \ldots S_N$ are independent with $S_i \sim \mathcal{G}(\sigma_i^2)$, then $\sum_{i=1}^N S_i \sim \mathcal{G}\left(\sum_{i=1}^N \sigma_i^2\right)$.

Combining these properties, we have (Lemma G.31)

$$S_i \sim \mathcal{G}(\sigma_i^2) \quad \text{independently} \quad \Rightarrow \quad \Pr\left(\left|\frac{1}{N} \sum_{i=1}^N S_i\right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{N^2 \epsilon^2}{2 \sum_{i=1}^N \sigma_i^2}\right)$$

(24)

An important class of sub-Gaussian variables is bounded scalars: if $X \in [a, b]$ then $X - E[X] \sim \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$ (Hoeffding’s lemma, G.27). This yields the following popular tail inequality.

**Corollary 7.1** (Hoeffding’s inequality). If $X_1 \ldots X_N \in [a, b]$ are iid with mean $\mu = E[X_i] \in \mathbb{R}$,

$$\Pr\left(\left|\frac{1}{N} \sum_{i=1}^N X_i - \mu\right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2N \epsilon^2}{(b-a)^2}\right)$$

(25)

**Proof.** By Hoeffding’s lemma, $X_i - \mu \sim \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$. We get the statement by plugging $\sigma_i^2 = \frac{(b-a)^2}{4}$ in (24). \qed

8 Cumulative Distribution Function

The cumulative distribution function (CDF) $\Phi : (-\infty, \infty) \rightarrow (0, 1)$ of the standard normal distribution is\(^4\)

$$\Phi(a) := \Pr(X \leq a) = \int_{-\infty}^{a} \mathcal{N}(0, 1)(x)dx$$

where $\Phi(0) = \frac{1}{2}$ (by symmetry) and $\Phi'(a) = \mathcal{N}(0, 1)(a)$ (by the fundamental theorem of calculus). One use of $\Phi$ is approximating $\text{sigmoid}(a) := (1 + e^{-a})^{-1}$. We find $\lambda$ so that the slope of $\Phi(\lambda a)$ is the same as that of $\text{sigmoid}(a)$ at 0. This yields (Lemma G.33)

$$\Phi\left(\sqrt{\frac{\pi}{8}} a\right) \approx \text{sigmoid}(a)$$

(26)

The quality of the approximation is visually apparent in the figure. Another useful property of $\Phi$ is that it is closed under a Gaussian expectation. For any $\lambda, \beta \in \mathbb{R}$ (Lemma G.34):

$$\mathbb{E}_{X \sim \mathcal{N}(\mu, \sigma^2)}[\Phi(\lambda X + \beta)] = \Phi\left(\frac{\lambda \mu + \beta}{\sqrt{1 + \lambda^2 \sigma^2}}\right)$$

(27)

(26) and (27) can be used together to derive an approximate closure of sigmoid under a Gaussian expectation:

$$\mathbb{E}_{X \sim \mathcal{N}(\mu, \sigma^2)}[\text{sigmoid}(X)] \approx \text{sigmoid}\left(1 + \frac{\pi \sigma^2}{8}\right)^{-1/2} \mu$$

(28)

\(^4\)For any distribution $f$ over $X \in \mathbb{R}^d$ the associated CDF is $F(a) := \Pr(X_1 \leq a_1 \land \cdots \land X_d \leq a_d)$. We focus on the one-dimensional standard normal because it is the most useful.
9 Gaussian Processes

A Gaussian process (GP) is a distribution over functions \( f : \mathcal{X} \to \mathbb{R} \) such that for any \( x \in \mathcal{X}^N \):

\[
f(x) \sim \mathcal{N}(0_N, k(x))
\]  

where \( k(x) \in \mathbb{R}^{N \times N} \) denotes the Gram matrix using a kernel \( k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \). We may “sample a function” by first building the giant Gram matrix \( k(\mathcal{X}) \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|} \) (pretending \( \mathcal{X} \) is finite) and drawing \( f \sim \mathcal{N}(0_{|\mathcal{X}|}, k(\mathcal{X})) \). Let \( x \in \mathcal{X}^N \) denote all training inputs and \( x_{\text{test}} \in \mathcal{X}^M \) all test inputs. We observe the training labels \( y = f(x) + \epsilon \) with Gaussian noises \( \epsilon \sim \mathcal{N}(0_N, \Sigma) \) where \( \Sigma > 0 \); the predictions are noiseless. We have

\[
\begin{bmatrix}
y \\
f(x_{\text{test}})
\end{bmatrix} \sim \mathcal{N}
\begin{bmatrix}
0_{N + M} \\
0_{N + M} + \Sigma
\end{bmatrix},
\begin{bmatrix}
k(x, x) + \Sigma & k(x, x_{\text{test}}) \\
k(x_{\text{test}}, x) & k(x_{\text{test}}, x_{\text{test}})
\end{bmatrix}
\]

By the Gaussian chain rule (7), we have the “predictive posterior” (implicitly conditioning on all inputs \( x, x_{\text{test}} \))

\[
f(x_{\text{test}}) \mid y \sim \mathcal{N}
\begin{bmatrix}
(k(x_{\text{test}}, x)(k(x) + \Sigma)^{-1} y, k(x_{\text{test}}, x)(k(x) + \Sigma)^{-1} k(x, x_{\text{test}})
\end{bmatrix}
\]

which generalizes Bayesian linear regression (86). We see that in the noiseless case \( \Sigma = 0_{N \times N} \), the prediction for the \( i \)-th training input \( x_i \) is \( f(x_i) \) with probability 1. The choice of the kernel can be guided by the marginal log-likelihood (MLL) \( \log \mathcal{N}(0_N, k(x) + \Sigma)(y) \). Below we visualize some samples of functions \( f : \mathbb{R} \to \mathbb{R} \) drawn from the prior (29) and the predictive posterior (31) (using an RBF kernel):

We can augment a GP over \( f : \mathcal{X} \to \mathbb{R} \) to define a distribution over all (binary) classifiers \( L : \mathcal{X} \to \{0, 1\} \) where \( L(x) = 1 \) with probability \( \text{sigmoid}(f(x)) \). We treat \( y = f(x) + \epsilon \in \mathbb{R}^N \) and \( f(x_{\text{test}}) \in \mathbb{R}^M \) in (30) as logits and assume the training and test labels \( L(x) \in \{0, 1\}^N \) and \( L(x_{\text{test}}) \in \{0, 1\}^M \) distributed as

\[
\begin{bmatrix}
L(x) \\
L(x_{\text{test}})
\end{bmatrix} \sim \text{Ber} \left( \text{sigmoid} \left( \begin{bmatrix} y \\ f(x_{\text{test}}) \end{bmatrix} \right) \right)
\]

To classify the test points it is sufficient to infer \( p_{\text{test}} \in (0, 1)^M \) where

\[
p_{\text{test}} = \mathbb{E}_{f_{\text{test}} \sim \text{posterior}(\cdot \mid L(x))} \left[ \text{sigmoid}(f_{\text{test}}) \right]
\]

Here, \( \text{posterior} (\cdot \mid L(x)) \) is a posterior distribution over the (noisy) training logits \( y \) conditioned on the observed labels. Conditioned on the training logits, the test logits are distributed as the predictive posterior (31). A closed-form approximation \( p_{\text{test}} \) of (32) is (Lemma G.35)

\[
[p_{\text{test}}]_j = \text{sigmoid} \left( \frac{1}{8} \left( 1 + \pi \left( k([x_{\text{test}}]_j) - k([x_{\text{test}}]_j, x)A^{-1}k(x, [x_{\text{test}}]_j) \right)^{-1/2} \right) \times k([x_{\text{test}}]_j, x)(L(x) - \text{sigmoid}(y^*)) \right)
\]

\( ^5 \) A GP may be viewed as a generalization of joint normality to infinitely many random variables. GPs do not express all functions. The choice of the kernel imposes certain functional limitations (e.g., RBF implies that we only consider smooth functions). We assume that the function values are mean-centered without loss of generality.
where \( y^* \in \mathbb{R}^N \) is the mode of the posterior and \( A = k(x) + \Sigma + \text{diag}(\text{sigmoid}(y^*) \odot (1 - \text{sigmoid}(y^*)))^{-1} \). The mode can be found by numerically optimizing the unnormalized log posterior which is convex (e.g., by Newton’s method). Below we visualize a ground-truth classifier (sky blue line), obtained by sampling a logit function from (29) (again using an RBF kernel) then squashing it by the sigmoid. Training labels (\( + = 1, - = 0 \)) are generated according to these probabilities on random points:

We see that classification has more problems with noise because the labels are discrete (even when the logits themselves are noiseless). The red lines show the model probabilities (33) using different RBF bandwidths; the first one uses the ground-truth value. The corresponding decision rules are shown in transparent green. The model is rather conservative to account for the noisy nature of the discrete labels. Nonetheless, we can force the model to fit the training data by a small bandwidth (second model).

### 9.1 Sparse GPs

A **sparse GP** (for regression) introduces a set of \( m \ll N \) inputs \( x_m \in \mathcal{X}^m \) whose labels are unknown, in addition to the usual training inputs \( x \in \mathcal{X}^N \) and their labels \( y \in \mathbb{R}^N \). It assumes that a latent \( z = (f_m, f) \in \mathbb{R}^m \times N \) is first drawn from the Gaussian process (29) on inputs \( (x_m, x) \in \mathcal{X}^{m+N} \), then the noisy labels \( y \sim \mathcal{N}(f, \Sigma) \) of \( x \) are observed. In this setting, we can derive the exact posterior \( p(f_m|y) \) and the MLL log \( p(y) \) (Lemma G.36):

\[
p(f_m|y) = \mathcal{N}(\Lambda(x_m)^{-1}k(x_m)\Sigma^{-1}y, \Lambda(x_m)^{-1}) \quad (f_m) \tag{34}
\]
\[
\log p(y) = \log \mathcal{N}(0_N, Q(x_m) + \Sigma)(y) - \frac{1}{2} \text{tr}(\Sigma^{-1}(k(x) - Q(x_m))) \quad (35)
\]

where \( \Lambda(x_m) = k(x_m)^{-1} + k(x_m)^{-1}k(x_m, x)\Sigma^{-1}k(x_m, x_m)k(x_m)^{-1} \) and \( Q(x_m) = k(x, x_m)k(x_m)^{-1}k(x_m, x) \). Since the additional inputs only change the prior, \( (35) \) is similar to the MLL in a standard GP (i.e., \( \log \mathcal{N}(0_N, k(x) + \Sigma)(y) \)), plus a regularization term that penalizes the difference between \( k(x) \) and \( Q(x_m) \). A sparse GP consists of two steps:

1. **(Training)** Since \( (35) \) is typically a differentiable function of \( x_m \) and kernel hyperparameters, we first perform gradient-based optimization to estimate their optimal values.

2. **(Inference)** Once the desired variables in \( (35) \) are optimized, given test points \( x_{\text{test}} \in \mathcal{X}^{M} \) we compute the posterior distribution over \( f(x_{\text{test}}) \in \mathbb{R}^{M} \) by marginalizing out \( f_m \) in \( y \rightarrow f_m \rightarrow f(x_{\text{test}}) \). Since \( y \rightarrow f_m \) is given by \( (34) \) and \( f_m \rightarrow f(x_{\text{test}}) \) by \( (31) \) (without the noise covariance), by (9)

\[
f(x_{\text{test}}|y) \sim \mathcal{N}\left(k(x_{\text{test}}, x_m)k(x_m)^{-1}\mu(x_m, y), \right.
\]
\[
\left. k(x_{\text{test}}) - k(x_{\text{test}}, x_m)k(x_m)^{-1}k(x_m, x_{\text{test}}) + k(x_{\text{test}}, x_m)(k(x_m)\Lambda(x_m)k(x_m))^{-1}k(x_m, x_{\text{test}}) \right)
\tag{36}
\]

A critical advantage of a sparse GP is that it can avoid the \( O(N^3) \) computational bottleneck of inverting the \( N \times N \) Gram matrix \( (31) \). Computing \( (35) \) can be done in \( O(Nm^2 + m^3) \) assuming a simple noise distribution (e.g., \( \Sigma = \sigma^2 I_{N \times N} \)). Computing \( (36) \) can be done in \( O(Mm^2) \). The details of how to invert the covariance matrix efficiently can be found in this note.
10 TODO: High-Dimensional Behavior

References


A Integration

A.1 Single-Variable

An antiderivative of \( f : \mathbb{R} \to \mathbb{R} \) is a function \( F : \mathbb{R} \to \mathbb{R} \) such that \( F' = f \). If \( F \) is an antiderivative, then so is \( F + C \) for any constant \( C \in \mathbb{R} \). For instance, \((1/3)x^3 + 42\) is an antiderivative of \( x^2 \).

The (definite) integral of \( f : \mathbb{R} \to \mathbb{R} \) over \([a, b]\) is a scalar \( \int_a^b f(x) \, dx \in \mathbb{R} \) that represents the signed area of \( f \) on \([a, b]\). The quantity \( f(x) \, dx \) is interpreted as the product of the function value and an infinitesimally small interval. There are different ways to formalize the area. The most common definition is the Riemann integral which partitions \([a, b]\) into intervals \([i\delta, (i + 1)\delta]\) of width \( \delta > 0 \) and define

\[
\int_a^b f(x) \, dx := \lim_{\delta \to 0} \sum_i f(x_i^{\delta})\delta
\]  

(37)

where \( x_i^{\delta} \in [i\delta, (i + 1)\delta] \). The finite sum \( \sum_i f(x_i^{\delta})\delta \) for a given width \( \delta \) is called a Riemann sum. Thus an integral is simply the limiting value of a Riemann sum (if it exists it is unique). A more general definition is a Lebesque integral which partitions the range of \( f \).

The fundamental theorem of calculus (FTC) allows us to evaluate integrals by antiderivatives: if \( F \) is any antiderivative of \( f \), then

\[
\int_a^b f(x) \, dx = F(x) \bigg|_a^b := F(b) - F(a)
\]  

(38)

For instance, the signed area under \( x^2 \) over \([-1, 1]\) is 2/3. Basic properties of integration include

\[
\int_a^b \alpha f(x) + \beta g(x) \, dx = \alpha \int_a^b f(x) \, dx + \beta \int_a^b g(x) \, dx
\]  

(linearity)

(39)

\[
\int_a^b f(g(x))g'(x) \, dx = \int_{g(a)}^{g(b)} f(u) \, du
\]  

(\(u\)-substitution)

(40)

\[
\int_a^b f(x)G(x) \, dx = F(x)G(x) \bigg|_a^b - \int_a^b F(x)g(x) \, dx
\]  

(integration by parts)

(41)

(Exercise: verify (40–41) using the chain rule and the product rule in differentiation.)

A.1.1 Substitution in practice

While (40) is the standard form of \( u \)-substitution, we often use it mechanically as follows. We wish to integrate \( f \) over the interval \( a < b \). We view \( f \) as a (hopefully simpler) function of \( u = g(x) \) where \( g : \mathbb{R} \to \mathbb{R} \) is invertible and differentiable with nonzero derivative over \((a, b)\). The infinitesimals are related as \( du = g'(x) \, dx \) by the chain rule, or equivalently \( dx = g'(g^{-1}(u))^{-1} \, du \). This yields a “plug-in” version of (40) where we substitute \( g(x) = u \) and \( dx = g'(g^{-1}(u))^{-1} \, du \),

\[
\int_a^b f(g(x)) \, dx = \int_{g(a)}^{g(b)} f(u)g'(g^{-1}(u))^{-1} \, du
\]  

(42)

For instance,

\[
\int_0^{\sqrt{2}} 2x \cos (x^2) \, dx = \int_0^{\sqrt{2}} 2\sqrt{u} \cos (u) \left( \frac{1}{2\sqrt{u}} \right) \, du = \int_0^{\sqrt{2}} \cos (u) \, du = \sin (u) \bigg|_0^{\sqrt{2}} = 1
\]

where \( 2x \cos (x^2) = 2\sqrt{u} \cos (u) \) with \( u = g(x) = x^2 \). Note that \( g \) is invertible on \((0, \sqrt{2})\) so that \( x = \sqrt{u} \); it is also differentiable with nonzero derivative \( g'(x) = 2x \). Writing \( dx = (2\sqrt{u})^{-1} \, du \), we cancel terms and are finally able to use FTC (38).
Orientation of region. Observe that

\[ 1 = \int_{-1}^{1} 1 \, dx = \int_{0}^{1} (-1) \, du = \int_{-1}^{0} (+1) \, du \]

The first equality is by FTC. The second equality is by (42) with \( f(x) = 1 \) and \( u = g(x) = -x \). The final equality is again by FTC, simply acknowledging that \((-x)\big|_{0}^{-1} = x\big|_{0}^{1} = 1\). More generally, when \( g'(x)^{-1} < 0 \) (i.e., \( u \) is moving in the opposite direction of \( x \)), we also change the “orientation of region” in integration (right-to-left instead of left-to-right). We can consider an alternative orientation-free formulation of \( u \)-substitution by always assuming integrating left-to-right. Let \( R \) denotes a region \( a < b \), then

\[ \int_{R} f(g(x)) \, dx = \int_{g(R)} f(u) |g'(g^{-1}(u))^{-1}| \, du \] (43)

where \( g(R) \) is the output region of \( g \) when applied to \( R \), integrated from a smaller value to a larger value. This formulation is useful because it generalizes to higher dimensions (45).

### A.2 Multi-Variable

The integral of \( f : \mathbb{R}^{d} \to \mathbb{R} \) over a region \( R \subseteq \mathbb{R}^{d} \) is a scalar \( \int_{R} f(x) \, dx \in \mathbb{R} \) that represents the signed hypervolume of \( f \) on \( R \). Evaluation of such an integral is generally challenging because the region may take complicated forms (high-dimensional curves).

We can greatly simplify the problem by restricting the region to be a hypercube \( R = [a, b] \) where \( a, b \in \mathbb{R}^{d} \) specify a \( d \)-dimensional bounding box \([a_1, b_1] \times \cdots \times [a_d, b_d]\) (potentially all of \( \mathbb{R}^{d} \)). A central tool in this setting is Fubini’s theorem, which states that

\[ \int_{[a, b]} f(x) \, dx = \int_{[a,b]}^{b_{\pi(d)}} \left( \cdots \left( \int_{a_{\pi(1)}}^{b_{\pi(1)}} f(x_1, \ldots, x_d) \, dx_{\pi(1)} \right) \cdots \right) \, dx_{\pi(d)} \]

where \( \pi \) is any permutation of \( \{1 \ldots d\} \). Thus we can evaluate a multi-variable integral by iteratively evaluating a single-variable integral in any order.

Many properties of integration carry over (like linearity), but some need to be generalized. One important generalization is multi-variable \( u \)-substitution. Let \( R \subseteq \mathbb{R}^{d} \) and \( g : R \to \mathbb{R}^{d} \) such that \( J_{g}(x) \in \mathbb{R}^{d \times d} \) (Jacobian of \( g \)) is nonzero for all \( x \in R \). Then

\[ \int_{R} f(g(x)) \, |\text{det}(J_{g}(x))| \, dx = \int_{g(R)} f(u) \, du \] (44)

Similar to the single-variable case, we often use substitution mechanically as follows. We integrate \( f \) over a region \( R \) by viewing it as a simpler function of \( u = g(x) \) where \( g : R \to \mathbb{R}^{d} \) is assumed to be invertible (i.e., \( \text{det}(J_{g}(x)) \neq 0 \)). The infinitesimals are related as \( du = |\text{det}(J_{g}(x))| \, dx \) or equivalently \( dx = |\text{det}(J_{g}(x))|^{-1} \, du \). This gives

\[ \int_{R} f(g(x)) \, dx = \int_{g(R)} f(u) \, |\text{det}(J_{g}(g^{-1}(u))]|^{-1} \, du \] (45)

where we “plug in” \( g(x) = u \) and \( dx = |\text{det}(J_{g}(g^{-1}(u)))]^{-1} \, du \). This strictly generalizes (43).

### A.2.1 Applications to probability

Let \( X \in \mathbb{R}^{d} \) be a random vector with distribution \( p_{X} \) supported on \( S \subseteq \mathbb{R}^{d} \) (i.e., \( p_{X}(x) \geq 0 \) and \( \int_{S} p_{X}(x) \, dx = 1 \)). The probability that \( X \) lies in a region \( R \subseteq S \) is

\[ \Pr(X \in R) = \int_{R} p_{X}(x) \, dx \]

Let \( t : S \to T \) be a smooth invertible function where \( T \subseteq \mathbb{R}^{d} \). Define a new random vector \( Y = t(X) \) supported on \( T \). We claim that \( Y \) has the distribution

\[ p_{Y}(y) = p_{X}(t^{-1}(y)) |\text{det}(J_{t^{-1}}(y))| \]

\[ \forall y \in T \] (46)
Equivalently,

\[ p_Y(t(x)) = p_X(x) |\det(J_{t^{-1}}(t(x)))| \quad \forall x \in S \tag{47} \]

**Proof sketch.** For any \( R \subseteq T \),

\[ \Pr(Y \in R) = \Pr(X \in t^{-1}(R)) = \int_{t^{-1}(R)} p_X(x)dx = \int_R p_X(t^{-1}(y)) |\det(J_{t^{-1}}(y))| dy \]

where the last equality applies (44) with \( g = t^{-1} \). This implies (46).

**B Continuous Entropy and KL Divergence**

We generalize results in Marsh (2013) to multivariate. The continuous/differential entropy of \( X \in \mathbb{R}^d \) with density \( p_X \) supported on \( S \subseteq \mathbb{R}^d \) is defined as\(^6\)

\[ H(X) := -\int_S p_X(x) \log p_X(x)dx \tag{48} \]

It is easily seen that entropy is additive for independent variables. That is, if \( X \in \mathbb{R}^d \) and \( Y \in \mathbb{R}^{d'} \) are independent then the entropy of \( Z = (X, Y) \in \mathbb{R}^{d+d'} \) is \( H(Z) = H(X) + H(Y) \).

- The uniform distribution \( u_{[a,b]}(x) := \frac{1}{b-a} \) over \( [a, b] \subset \mathbb{R} \) has entropy
  \[ H(X) = \int_a^b \frac{1}{b-a} \log(b-a)dx = \log(b-a) \tag{49} \]

- The Gaussian distribution \( \mathcal{N}(\mu, \Sigma) \) over \( \mathbb{R}^d \) has entropy (Corollary G.7)
  \[ H(X) = \frac{1}{2} \log((2\pi e)^d |\det(\Sigma)|) \]

- The exponential distribution \( e_{\lambda}(x) := \lambda \exp(-\lambda x) \) over \( [0, \infty) \) with parameter \( \lambda > 0 \) has entropy (Lemma G.5)
  \[ H(X) = 1 - \log \lambda \tag{50} \]

Unfortunately, continuous entropy suffers from various shortcomings (reviewed in Section B.1), most notably negativity (e.g., (49) is negative if \( b - a < 1 \), (50) is negative if \( \lambda > e \)). On the other hand, let \( q_X \) be another density of \( X \) with support \( S \). Define the continuous KL divergence (aka. relative entropy) between \( p_X \) and \( q_X \) as

\[ \text{KL}(p_X, q_X) := \int_S p_X(x) \log \frac{p_X(x)}{q_X(x)} dx \tag{51} \]

Continuous KL divergence is nonnegative:
\[
\text{KL}(p_X, q_X) = \mathbb{E}_{x \sim p_X} \left[ \log \frac{p_X(x)}{q_X(x)} \right] \\
= \mathbb{E}_{x \sim p_X} \left[ -\log \frac{q_X(x)}{p_X(x)} \right] \\
\geq -\log \left( \mathbb{E}_{x \sim p_X} \left[ \frac{q_X(x)}{p_X(x)} \right] \right) \\
= -\log \left( \int_S p_X(x) \frac{q_X(x)}{p_X(x)} dx \right) \\
= -\log \left( \int_S q_X(x)dx \right) = 0
\]

where \( \text{KL}(p_X, q_X) = 0 \) iff \( p_X = q_X \) almost everywhere. This has useful implications.

\(^6\)We use the term “density” in this section to distinguish continuous vs discrete variables.
• The cross entropy between \( p_X \) and \( q_X \) upper bounds the entropy of \( p_X \),
\[
H(p_X, q_X) := H(p_X) + \text{KL}(p_X, q_X) \geq H(p_X)
\] (52)

• Mutual information is nonnegative,
\[
I(X, Y) := \text{KL}(p_{XY}, p_X p_Y) \geq 0
\] (53)

The cross entropy upper bound can be used to derive various maximum entropy densities.

**Theorem B.1.**
\[
\mathcal{N}(\mu, \Sigma) \in \arg \max_{p_X : \mathbb{E}[X] = \mu, \text{Var}(X) = \Sigma} H(p_X) \tag{54}
\]
\[
u_{[a, b]} \in \arg \max_{p_X : \text{Support}(p_X) = [a, b]} H(p_X) \tag{55}
\]
\[
\epsilon_\lambda \in \arg \max_{p_X : \text{Support}(p_X) = \mathbb{R}^d_+, \mathbb{E}[X] = \lambda^{-1}} H(p_X) \tag{56}
\]

where \( u_{[a, b]} \) denotes the uniform distribution over \([a, b] \subset \mathbb{R}^d\) and \( \epsilon_\lambda \) denotes the product exponential density over \( \mathbb{R}^d_+ \) with \( \lambda > 0 \).

**Proof.** (54): Let \( p_X \) with mean \( \mu \in \mathbb{R}^d \) and covariance \( \Sigma \succ 0 \). Then
\[
H(p_X, \mathcal{N}(\mu, \Sigma)) = \int_{\mathbb{R}^d} p_X(x) \left( \frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) + \frac{1}{2} \log((2\pi)^d \det(\Sigma)) \right)
\]
\[
= \frac{1}{2} \mathbb{E}_{x \sim p_X} [(x - \mu)^\top \Sigma^{-1} (x - \mu)] + \frac{1}{2} \log((2\pi)^d \det(\Sigma))
\]
\[
= \frac{d}{2} + \frac{1}{2} \log((2\pi)^d \det(\Sigma))
\]
\[
= \frac{1}{2} \log((2\pi e)^d \det(\Sigma)) = H(\mathcal{N}(\mu, \Sigma)) \geq H(p_X)
\]

(55): Assume \( d = 1 \). Given any \( p_X \) with support \([a, b]\) we have
\[
H(p_X, u_X) = \int_a^b p_X(x) \log(b - a)dx = \log(b - a) = H(u_{[a, b]}) \geq H(p_X)
\]

The statement holds for \( d > 1 \) since each dimension is independently optimized.

(56): Assume \( d = 1 \). Given any \( p_X \) with support \([0, \infty)\) and mean \( \lambda^{-1} > 0 \) we have
\[
H(p_X, e_\lambda) = \int_0^\infty p_X(x) (\lambda x - \log \lambda)dx = \lambda \mathbb{E}_{x \sim p_X} [x] - \log \lambda = 1 - \log \lambda = H(e_\lambda) \geq H(p_X, e_\lambda)
\]

The statement holds for \( d > 1 \) since each dimension is independently optimized. \(\square\)

**B.1 Shortcomings of Continuous Entropy**

**B.1.1 Inconsistency with Shannon entropy**

The Shannon entropy of discrete \( X \in \{x_1 \ldots x_n\} \) with distribution \( p_X \) is
\[
H(X) := -\sum_{i=1}^n p_X(x_i) \log p_X(x_i)
\] (57)

This definition was derived by Shannon as a solution that satisfies axioms of information (regarding monotonicity, non-negativity, zero information, and independence). (48) appears to be a natural continuous extension of (57) in the sense that both are \( \mathbb{E}_{x \sim p_X} [-\log p_X(x)] \), but it fails to satisfy the axioms (e.g., it can be negative). One way to
better understand why is to show that (48) is inconsistent with (57) in the limit. Assume $d = 1$ and let $p_X$ be a density supported on $[a, b]$. By definition
\[
\int_a^b p_X(x)dx = \lim_{\delta \to 0} \sum_i p_X(x_i^\delta)\delta = 1
\] (58)
where $\sum_i p_X(x_i^\delta)\delta$ is a finite Riemann sum of width $\delta > 0$. Thus we can cast the density $p_X$ as an increasingly fine-grained discrete distribution with probabilities $p_X(x_i^\delta)\delta$ as $\delta \to 0$. Note that each value of $\delta > 0$ yields a discrete distribution with a well-defined Shannon entropy. This Shannon entropy, in the limit, is
\[
\lim_{\delta \to 0} \left( -\sum_i (p_X(x_i^\delta)\delta) \log(p_X(x_i^\delta)\delta) \right) = -\lim_{\delta \to 0} \sum_i (p_X(x_i^\delta) \log p_X(x_i^\delta))\delta - \lim_{\delta \to 0} \sum_i p_X(x_i^\delta)\delta \log \delta
\]
\[
= -\int_a^b p_X(x) \log p_X(x)dx - \lim_{\delta \to 0} \sum_i p_X(x_i^\delta)\delta \log \delta
\]
\[
= H(X) - \left( \lim_{\delta \to 0} \sum_i p_X(x_i^\delta)\delta \right) \left( \lim_{\delta \to 0} \log \delta \right)
\] (59)
\[
= H(X) + \infty
\] (60)
where (59) follows from the generalized product rule of limits using (58). So the limiting Shannon entropy diverges from the continuous entropy by an infinite offset.

**B.1.2 Variability under change of coordinates**

A good measure of information should not depend on the representation of samples from a distribution. For instance, let $p_X$ be a distribution over finitely many circles, each of which can be specified by its radius or area. Clearly, the Shannon entropy of the circle is the same regardless of the representation. Now let $p_X$ be a density over all circles. The continuous entropy of the circle under the radius representation is different from that under the area representation. A general statement that implies this result is given below.

**Lemma B.2.** Let $X \in \mathbb{R}^d$ with density $p_X$ supported on $S$. For any invertible mapping $t$ on $S$,
\[
H(t(X)) = H(X) - \mathbb{E}_{x \sim p_X} [\log |\det(J_{t^{-1}}(t(x)))|]
\]

**Proof.**
\[
H(t(X)) = -\int_S p_X(x) \log p_X(t(x))dx
\]
\[
= -\int_S p_X(x) \log p_X(x)dx - \int_S p_X(x) \log |\det(J_{t^{-1}}(t(x)))|dx
\] (by (47))
\[
= H(X) - \mathbb{E}_{x \sim p_X} [\log |\det(J_{t^{-1}}(t(x)))|]
\]

\[\square\]

**Corollary B.3.** For any invertible $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$,
\[
H(AX + b) = H(X) - \log |\det(A^{-1})|
\] (61)

**Corollary B.4.** For $\alpha > 0$,
\[
H(\alpha X) = H(X) + d \log \alpha
\]

---

7Assume $\lim_{x \to a} f(x) \neq 0$. If $g(x)$ does not oscillate around $a$,
\[
\lim_{x \to a} f(x)g(x) = \lim_{x \to a} f(x) \lim_{y \to a} g(y)
\]
If $g(x)$ oscillates around $a$, then so does $f(x)g(x)$.
We use the same substitution in the proof of Lemma G.4. Let $\Sigma = \text{eigendecomposition}$. Let $u \in J$.

Proof. $H(\alpha X) = H(X) - \log |\det(\alpha^{-1}I_{d\times d})|$ (by (61))

$= H(X) - \log |\alpha^{-d}|$

$= H(X) - \log \alpha$

$= H(X) + d \log \alpha$ (since $\alpha > 0$)

Corollary B.4 states that we can vacuously increase the continuous entropy of $X \in \mathbb{R}^d$ to infinity by multiplying each value with a scalar $\alpha$ as we take $\alpha \to \infty$.

C Moment-Generating Function

Let $X \in \mathbb{R}^d$ denote a random vector with distribution $p_X$. The moment-generating function (MGF) of $X$ is a real-valued positive mapping $M_X : \mathbb{R}^d \to (0, \infty)$ defined as

$$M_X(t) := \mathbb{E}_{x \sim p_X} [\exp (t^T x)]$$

(62)

Not every distribution has a corresponding MGF (because (62) may diverge). But a classical result in probability theory is that an MGF uniquely determines a probability distribution. More formally, let $X, Y \in \mathbb{R}^d$ be random vectors with distributions $p_X, p_Y$ with well-defined MGFs $M_X, M_Y$. Then $p_X = p_Y$ iff $M_X = M_Y$. Thus an MGF is an alternative characterization of a random variable.

What makes $M_X$ special is obviously the exponential function. Since $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$,

$$M_X(t) = 1 + t^T \mathbb{E}[X] + \frac{1}{2} t^T \mathbb{E}[XX^T] t + \cdots$$

so that $\nabla^n M_X(0)$ is the $n$-th moment of $p_X$ (hence the name).

Lemma C.1. Let $X \sim \mathcal{N}(\mu, \Sigma)$. Then

$$M_X(t) = \exp \left( t^T \mu + \frac{1}{2} t^T \Sigma t \right)$$

Proof. We use the same substitution in the proof of Lemma G.4. Let $\Sigma = U\Lambda U^\top$ denote an orthonormal eigendecomposition. Let $u = g(x)$ where $g(x) = \Lambda^{-1/2}U^\top (x - \mu)$, which implies $x = U\Lambda^{1/2}u + \mu$. Thus $|\det(J_g(x))| = |\det(\Lambda^{-1/2}U^\top)| = |\det(\Lambda)^{-1/2}|$, so we have the infinitesimal $dx = \sqrt{|\det(\Lambda)|} du$. Then

$$\int_{\mathbb{R}^d} \frac{1}{(2\pi)^d \sqrt{|\det(\Sigma)|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \exp(t^T x) dx$$

$$= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d \sqrt{|\det(\Lambda)|}} \exp \left( -\frac{1}{2} u^T u \right) \exp(t^T U\Lambda^{1/2}u + t^T \mu) du$$

$$= \exp(t^T \mu) \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d} \exp \left( -\frac{1}{2} u^T u + t^T U\Lambda^{1/2}u \right) du$$

$$= \exp(t^T \mu) \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d} \exp \left( -\frac{1}{2} \left| u - U\Lambda^{1/2}t \right|^2 \right) du$$

$$= \exp \left( t^T \mu + \frac{1}{2} t^T \Sigma t \right) \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d} \exp \left( -\frac{1}{2} \left| u - U\Lambda^{1/2}t \right|^2 \right) du$$

$$= \exp \left( t^T \mu + \frac{1}{2} t^T \Sigma t \right)$$
The first two moments of the Gaussian MGF are
\[ h(t) := t^T \mu + \frac{1}{2} t^T \Sigma t \]
\[ \nabla M_X(t) = \exp(h(t)) (\mu + \Sigma t) \]
\[ \nabla^2 M_X(t) = \exp(h(t)) \Sigma + \exp(h(t)) (\mu + \Sigma t)^T (\mu + \Sigma t) \]
which imply that the mean and the covariance matrix of \( X \) are \( \mu \) and \( \Sigma \).

An interesting consequence of Lemma C.1 is that a point-mass density can be viewed as a degenerate Gaussian distribution with zero variance. That is, if \( X \in \mathbb{R}^d \) takes value \( a \in \mathbb{R}^d \) with probability 1, then \( M_X(t) = \exp(a^T t) \), which is equal to the Gaussian MGF with \( \Sigma = 0 \).

One application of MGF is showing that a linear transformation of a Gaussian random variable is also Gaussian. Note that the MGF of a linear transformation of \( X \) is generally
\[ M_{AX+b}(t) = \mathbb{E}_{x \sim p_X} \left[ \exp \left( (t^T A)x \right) \right] = \exp \left( (t^T b) M_X(A^T t) \right) \]

**Lemma C.2.** Let \( X \sim \mathcal{N}(\mu, \Sigma) \). Let \( A \in \mathbb{R}^{d'} \times d \) and \( b \in \mathbb{R}^d \) where \( d' \leq d \) and \( A \) has full rank. Then \( AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T) \).

**Proof.** For any \( t \in \mathbb{R}^d \),
\[ M_{AX+b}(t) = \exp \left( (t^T b) M_X(A^T t) \right) \]
\[ = \exp \left( (t^T b) \left( t^T A\mu + \frac{1}{2} t^T A\Sigma A^T t \right) \right) \]
\[ = \exp \left( t^T (A\mu + b) + \frac{1}{2} t^T A\Sigma A^T t \right) \]
The last term is the MGF of a random variable with distribution \( \mathcal{N}(A\mu + b, A\Sigma A^T) \) where \( A\Sigma A^T \succ 0 \). The statement follows from the one-to-one correspondence between MGFs and distributions.

**C.1 Cumulant-Generating Function**

The log MGF \( \psi_X(t) := \log \mathbb{E}[e^{t^T X}] \) is called the **cumulant-generating function (CGF)** of \( X \). We see that it is the (convex) log-partition function of \( t \)-tilted \( X_t \) distributed as (Appendix D)

\[ p_{X_t}(x) = \frac{e^{t^T x} p_X(x)}{\mathbb{E}[e^{t^T X}]} \]

We call \( \nabla^{(n)} \psi_X(t) \) the \( n \)-th **cumulant** of \( X \). From (73–74), we have
\[ \nabla \psi_X(t) = \mathbb{E}[X_t] \]
\[ \nabla^2 \psi_X(t) = \text{Cov}(X_t) \]
In particular,
\[ \nabla \psi_X(0_d) = \mathbb{E}[X] \]
\[ \nabla^2 \psi_X(0_d) = \text{Cov}(X) \]
This fact is used in Hoeffding’s lemma which bounds the CGF of a bounded scalar random variable by using Taylor’s approximation of the CGF around 0 and then bounding the mean/variance of that variable (Lemma G.27).

**Gaussian cumulants.** The CGF of \( X \sim \mathcal{N}(\mu, \Sigma) \) is \( \psi_X(t) = \mu^T t + \frac{1}{2} t^T \Sigma t \), so
\[ \nabla \psi_X(t) = \mu^T t + \frac{1}{2} t^T \Sigma t \]
\[ \nabla^2 \psi_X(t) = \Sigma \]
which is consistent with the fact that \( X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma) \) (Lemma G.26). The corresponding Legendre transform \( \psi^*_X(t) := \sup_{\lambda \in \mathbb{R}^d} \lambda^T t - \psi_X(\lambda) \) of \( \psi_X \) is (Lemma G.28)
\[ \psi^*_X(t) = \frac{1}{2} (t - \mu)^T \Sigma^{-1} (t - \mu) \]
D Exponential Family

D.1 Exponential Tilting

Given any “base” distribution $p$ over $\mathbb{R}^d$, we can generate a set of distributions $q_{p,\tau,\theta}$ by

$$q_{p,\tau,\theta}(x) := \frac{e^{\theta^\top \tau(x)} p(x)}{E_{x' \sim p}[e^{\theta^\top \tau(x')}]} \tag{71}$$

for any $\tau : \mathbb{R}^d \to \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ such that $E_{x' \sim p}[e^{\theta^\top \tau(x')}]$ exists. Note that

- $q_{p,\tau,\theta}$ is nonnegative and sums/integrates to 1.
- $q_{p,\tau,\theta}$ has the same support as $p$.
- $q_{p,\tau,\theta}$ assigns a weight $e^{\theta^\top \tau(x)}$ on the probability of $x$, changing the tails of $p$.
- $q_{p,\tau,0_m} = p$.

This technique is called exponential tilting of $p$. We can rewrite (71) as

$$q_{p,\tau,\theta}(x) = p(x) \exp (\theta^\top \tau(x) - B_{p,\tau}(\theta)) \tag{72}$$

where the log-partition function $B_{p,\tau}(\theta) := \log E_{x \sim p}[e^{\theta^\top \tau(x)}]$ normalizes $q_{p,\tau,\theta}$. We note several properties:

- $B_{p,\tau}(\theta)$ is convex (Lemma G.21).
- $\tau$ is a sufficient statistic for $\theta$ (Theorem G.20).
- Differentiating $B_{p,\tau}(\theta)$ generates the cumulants of $\tau(x)$ over $x \sim q_{p,\tau,\theta}$, for instance (Lemma G.22)

$$\nabla B_{p,\tau}(\theta) = E_{x \sim q_{p,\tau,\theta}} [\tau(x)] \tag{73}$$

$$\nabla^2 B_{p,\tau}(\theta) = \text{Cov}_{x \sim q_{p,\tau,\theta}} (\tau(x)) \tag{74}$$

In particular, $\nabla B_{p,\tau}(0_m) = E_{x \sim p} [\tau(x)]$ and $\nabla^2 B_{p,\tau}(0_m) = \text{Cov}_{x \sim p} (\tau(x))$.

- Aside: (74) implies that $B_{p,\tau}(\theta)$ is convex since $\nabla^2 B_{p,\tau}(\theta) \succeq 0$.

Exponential tilting often preserves the distribution family. For instance, if $X \sim \mathcal{N}(\mu, \Sigma)$ and $X_t$ is the $t$-tilted $X$ with $t \in \mathbb{R}^d$ ($\tau(x) = x$), then $X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma)$ (Lemma G.26).

D.2 Unnormalized Form

More generally, we may consider any nonnegative function $h : \mathbb{R}^d \to (0, \infty)$ (“base measure”) and define

$$q_{h,\tau,\theta}(x) = \frac{\exp (\theta^\top \tau(x)) h(x)}{\int_{x \in \mathbb{R}^d} \exp (\theta^\top \tau(x)) h(x) dx} \tag{75}$$

for any $\tau : \mathbb{R}^d \to \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ such that $\int_{x \in \mathbb{R}^d} \exp (\theta^\top \tau(x)) h(x) dx$ exists. We can rewrite (75) as

$$q_{h,\tau,\theta}(x) = h(x) \exp (\theta^\top \tau(x) - A_{h,\tau}(\theta)) \tag{76}$$

where $A_{h,\tau}(\theta) := \log (\int_{x \in \mathbb{R}^d} h(x) \exp (\theta^\top \tau(x)) dx)$ and $\tau$ is again a sufficient statistic for $\theta$. Clearly, exponential tilting is a special case where the base measure is normalized. However, (75) is strictly more general since it allows for $h$ such that $\int h(x) dx$ diverges. It is easy to check that the previous properties hold without a normalized base measure, specifically:

- Differentiating $A_{h,\tau}(\theta)$ generates the cumulants of $\tau(x)$ over $x \sim q_{h,\tau,\theta}$, in particular

$$\nabla A_{h,\tau}(\theta) = E_{x \sim q_{h,\tau,\theta}} [\tau(x)] \tag{77}$$

$$\nabla^2 A_{h,\tau}(\theta) = \text{Cov}_{x \sim q_{h,\tau,\theta}} (\tau(x)) \tag{78}$$

- (78) implies that $A_{h,\tau}(\theta)$ is convex.

A set of distributions that can be expressed in the form (76) is called an exponential family. $\theta \in \mathbb{R}^m$ is called its natural parameter. Note that there are many exponential families. For instance, the set of all normal distributions is one exponential family. The set of all categorical distributions is another exponential family.
D.2.1 Discussions

CGF. The CGF \( \psi_{\tau(X)}(t) = \log \mathbb{E}[e^{t^\top \tau(X)}] \) of \( \tau(x) \) takes the form (Lemma G.24):

\[
\psi_{\tau(X)}(t) = A_{\theta, \tau}(\theta + t) - A_{\theta, \tau}(\theta)
\]

(79)

where we see \( \nabla^{(n)} \psi_{\tau(X)}(0,m) = \nabla^{(n)} A_{\theta, \tau}(\theta) \); this is consistent with the fact that in an exponential family, the log-partition function generates cumulants.

Conjugate prior. In Bayesian probability theory, a prior over the parameter of a distribution is called a conjugate prior if the implied posterior over the parameter conditioning on a sample from the distribution is in the same distribution family that the prior is in. For an exponential family, we can define a prior

\[
\pi_{h, \tau}(\theta; \alpha, \beta) = \frac{1}{Z_{h, \tau}(\alpha, \beta)} \exp (\theta^\top \alpha - \beta A_{h, \tau}(\theta))
\]

(80)

for any “pseudo-counts” \( \alpha \in \mathbb{R}^m \) and \( \beta \in \mathbb{R} \) such that \( Z_{h, \tau}(\alpha, \beta) = \int_{\theta \in \mathbb{R}^m} \exp (\theta^\top \alpha - \beta A_{h, \tau}(\theta)) \, d\theta \) exists. Then the posterior over \( \theta \) given \( x \sim q_{h, \tau, \theta} \) is given by (Lemma G.25)

\[
k_{h, \tau}(\theta|x; \alpha, \beta) = \pi_{h, \tau}(\theta; \tau(x) + \alpha, 1 + \beta)
\]

(81)

thus (80) is a conjugate prior.

Identifying an exponential family. To check if a set of distributions \( \{p(x; \theta)\}_\theta \) is an exponential family, it is sufficient to propose any \( h(x) \geq 0 \), a transformation of \( \theta \) into natural parameter form \( \theta = g(\bar{\theta}) \in \mathbb{R}^m \) and \( x \) into sufficient statistic form \( \tau(x) \in \mathbb{R}^m \), and some function \( A_{h, \tau}(\theta) \), such that it can be written as (76):

\[
p(x; \bar{\theta}) = q_{h, \tau, \theta}(x) = h(x) \exp (\theta^\top \tau(x) - A_{h, \tau}(\theta))
\]

In particular, we do not need to explicitly calculate \( A_{h, \tau}(\theta) = \log (\int_{x \in \mathbb{R}^d} h(x) \exp (\theta^\top \tau(x)) \, dx) \) since the normalization of \( p(x; \bar{\theta}) \) enforces it (and guarantees its existence).

Non-unique parameterization. An exponential family has infinitely many equivalent parameterizations:

\[
q_{h, \tau, \theta}(x) = q_{\theta, h, u \odot \tau, \text{inv}(u) \odot \theta}(x) \quad \quad \forall \theta \in \mathbb{R}^m, u \in (\mathbb{R} \setminus \{0\})^m
\]

where \( \odot \) is the elementwise multiplication and \( \text{inv}(u) \) is the elementwise inverse of vector \( u \). It is often clear what a natural parameterization is (e.g., choose \( u \) that makes \( \tau(x) \) as simple as possible).

Limitations. A dizzying array of distributions are exponential families, including the normal (Lemma G.23), categorical, exponential, geometric, Bernoulli, Poisson, beta, and many others. But there are certain properties that an exponential family cannot capture. First, the form

\[
h(x) \exp (\theta^\top \tau(x) - A_{h, \tau}(\theta))
\]

implies that the support of this distribution cannot depend on the parameter \( \theta \). This rules out distributions like a uniform distribution on \([a, b] \subset \mathbb{R}\) whose support depends on the parameters \( a, b \). Second, some distributions simply cannot be expressed using an inner product between the input and the parameter, for instance the Laplace distribution

\[
\text{Laplace}(\mu, b)(x) = \frac{1}{2b} \exp \left( -\frac{|x - \mu|}{b} \right)
\]

Third, an exponential family necessarily has a well-defined MGF by (79), so it rules out distributions without an MGF such as the Cauchy distribution.
D.3 Tweedie’s Formula

Lemma D.1. Define the generative process over $t, x \in \mathbb{R}^d$:

$$t \sim p_T$$

$$x \mid t \sim p_{X \mid T}(\cdot \mid t)$$

where $b$ is some base distribution (i.e., we are exponential tilting it by $t$). Let

$$m(x) = \int_{t \in \mathbb{R}^d} p_T(t)p_{X \mid T}(x \mid t) dt$$

denote the marginal distribution over $x$. Define $l(x) := \log m(x)$ and $l_0(x) := \log b(x)$. We have

$$t \mid x \sim \text{Unk}(\nabla l(x) - \nabla l_0(x), \nabla^2 l(x) - \nabla^2 l_0(x))$$

Proof. By Bayes’ rule, the posterior over $t$ given $x$ is

$$p_{T \mid X}(t \mid x) = \frac{p_T(t)p_{X \mid T}(x \mid t)}{m(x)} = p_T(t)\mathbb{E}_{x \sim b}[e^{t^\top x}] \exp \left(t^\top x - \lambda(x)\right)$$

where $\lambda(x) = \log \frac{m(x)}{p_T(x)}$. This is an exponential family (76) with base measure $h(t) = p_T(t)\mathbb{E}_{x \sim b}[e^{t^\top x}]$, natural parameter $x \in \mathbb{R}^d$, sufficient statistic $t \in \mathbb{R}^d$, and the CGF $\lambda(x)$. By the usual property of the CFG, the mean $\mu$ and the covariance $\Sigma$ of $t \sim p_{T \mid X} (\cdot \mid x)$ is given by

$$\mu = \nabla \lambda(x) = \nabla l(x) - \nabla l_0(x)$$
$$\Sigma = \nabla^2 \lambda(x) = \nabla^2 l(x) - \nabla^2 l_0(x)$$

Lemma D.2 (Tweedie’s formula). Pick any $\Sigma > 0$. Define the generative process over $\mu, x \in \mathbb{R}^d$:

$$\mu \sim g$$

$$x \mid \mu \sim \mathcal{N}(\mu, \Sigma)$$

Let $m(x) = \int_{\mu \in \mathbb{R}^d} g(\mu)\mathcal{N}(\mu, \Sigma)(x)d\mu$ denote the marginal distribution over $x$. Define $l(x) := \log m(x)$. Then

$$\mu \mid x \sim \text{Unk}(x + \Sigma \nabla l(x), \Sigma(I_{d \times d} + \nabla^2 l(x)\Sigma))$$

Proof. We can view $x \mid \mu \sim \mathcal{N}(\mu, \Sigma) = \mathcal{N}(0_d + \Sigma t, \Sigma)$ as an exponential tilting of the base distribution $b(x) = \mathcal{N}(0_d, \Sigma)(x)$ by $t = \Sigma^{-1} \mu$ (Lemma G.26). Let $l_0(x) = \log b(x)$ and note that $\nabla l_0(x) = \nabla \left(-\frac{1}{2} x^\top \Sigma^{-1} x\right) = -\Sigma^{-1} x$. Lemma D.1 states that

$$t \mid x \sim \text{Unk}(\nabla l(x) - \nabla l_0(x), \nabla^2 l(x) - \nabla^2 l_0(x))$$

$$= \text{Unk}(\Sigma^{-1} x + \nabla l(x), \Sigma^{-1} + \nabla^2 l(x))$$

Thus $\mu = \Sigma t$ conditioned on $x$ is distributed as

$$\mu \mid x \sim \text{Unk}(x + \Sigma \nabla l(x), \Sigma + \Sigma \nabla^2 l(x)\Sigma)$$
E  Laplace Approximation

Let \( p_Z \) denote a prior over \( Z \in \mathbb{R}^d \) and \( p_{X|Z} \) a likelihood over \( X \) given \( Z \). Conditioned on \( X = x \), the Laplace approximation approximates the true posterior \( p_{Z|X}(z|x) \propto p_{X|Z}(x|z) \times p_Z(z) \) by a Gaussian:

\[
p_{Z|X}(z|x) \approx \mathcal{N}(z^*, -H_x(z^*)^{-1})
\]

where \( l_x(z) = \log p_{Z|X}(z|x) \) is the log posterior, \( z^* = \arg \max_{z \in \mathbb{R}^d} l_x(z) \), and \( H_x : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d} \) is the Hessian of \( l_x \) (which is assumed to exist). The approximation reflects the idea that the posterior is “pointy” around the mode and can be directly derived by the second-order Taylor approximation of \( l_x \) around \( z^* \),

\[
l_x(z^*) \approx \frac{1}{2}(z - z^*)^\top H_x(z^*)(z - z^*) + \text{constant}
\]

which, when normalized, becomes the distribution \( \mathcal{N}(z^*, -H_x(z^*)^{-1}) \).

F  Linear Regression

Let \( X = (x_1 \ldots x_N) \in \mathbb{R}^{N \times d} \) denote \( N \) input vectors in matrix form, paired with continuous labels \( y = (y_1 \ldots y_N) \in \mathbb{R}^N \). In “generalized” least squares, we assume \( y \sim \mathcal{N}(Xw^*, \Sigma) \) for some unknown \( w^* \in \mathbb{R}^d \) and known positive-definite \( \Sigma \in \mathbb{R}^{N \times N} \) (i.e., \( y_i = w^* \cdot x_i + \epsilon_i \) where \( \epsilon_1, \ldots, \epsilon_N \sim \mathcal{N}(0, \Sigma) \)). The hypothesis class is the family of conditional distributions \( \mathcal{N}(Xw, \Sigma) \) over \( \mathbb{R}^N \) indexed by \( w \in \mathbb{R}^d \). The maximum-likelihood estimator (MLE) with an \( l_2 \) regularization coefficient \( \lambda > 0 \) is

\[
\hat{w} = \arg \max_{w \in \mathbb{R}^d} \log(\mathcal{N}(Xw, \Sigma)(y)) - \frac{\lambda}{2} ||w||^2
\]

\[
= (X^\top \Sigma^{-1}X + \lambda I_{d \times d})^{-1}X^\top \Sigma^{-1}y
\]

\[
= X^\top (XX^\top + \lambda \Sigma)^{-1}y
\]

(83)

where (83) uses the matrix identity \((B^\top R^{-1}B + P^{-1})^{-1}B^\top R^{-1} = PB^\top (BPB^\top + R)^{-1}\) (Welling, 2013). Given a test point \( x_{test} \in \mathbb{R}^d \), the “true” label is produced by \( y_{test} = w^* \cdot x_{test} + \epsilon_{test} \) where \( \epsilon_{test} \sim \mathcal{N}(0, \nu_{test}) \). However, we predict \( \hat{y} = \hat{w}^\top x_{test} \), plugging in \( \hat{w} \) in place of \( w^* \) and assuming there is no test noise. Note that

\[
\hat{y} = y^\top (XX^\top + \lambda \Sigma)^{-1}Xx_{test}
\]

is both (1) a linear combination of the training labels, and (2) a linear combination of the dot products between the test point and training inputs. The latter view admits the kernel trick: compute the Gram matrix \( G \in \mathbb{R}^{N \times N} \) where \( G_{ij} = k(x_i, x_j) \) for a chosen kernel \( k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \) (i.e., representing \( \langle \phi(x_i), \phi(x_j) \rangle \) for some feature function \( \phi \)), compute \( \alpha = (G + \lambda \Sigma)^{-1}y \in \mathbb{R}^N \), and predict \( \hat{y} = \sum_{i=1}^{N} \alpha_i k(x_i, x_{test}) \). Note that regularization is necessary to obtain the kernelized version (this is why it is called kernel ridge regression) and to invoke the representer theorem.

F.1  Baysian Linear Regression

We assume \( w \sim \mathcal{N}(0_d, \Omega) \) for some known positive-definite \( \Omega \in \mathbb{R}^{d \times d} \). We now have a joint distribution \( p(w, y|X) = \mathcal{N}(0_d, \Omega)(w) \times \mathcal{N}(Xw, \Sigma)(y) \). By Gaussian Bayes’ rule (Section 3.3), the associated marginal and posterior distributions are

\[
p(y|X) = \mathcal{N}
\]

(0_N, \Sigma + XX^\top})(y)

\[
\pi(w|X, y) = \mathcal{N}
\]

(\Omega XX^\top (XX^\top + \Sigma)^{-1}y, \Omega - \Omega XX^\top (XX^\top + \Sigma)^{-1}XX^\top \Omega)(w)

(84)

where (84) uses the Woodbury identity \((A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1} \) (see (157) of The Matrix Cookbook). We see that the ridge regressor (83) corresponds to the mode (also mean) of the posterior (84) using \( \Omega = \lambda^{-1}I_{d \times d} \) (aka. the MAP estimate). Instead of using a single point, we can incorporate all of the posterior by considering the “predictive posterior”:

\[
p(\hat{y}|X, y, x_{test}) = \mathbb{E}_{w \sim \pi(w|X, y)} \left[ \mathcal{N}(w^\top x_{test}, \nu_{test})(\hat{y}) \right]
\]

\[
= \mathcal{N}
\]

(\hat{x}_{test}^\top \Omega XX^\top (XX^\top + \Sigma)^{-1}y, \nu_{test} + \hat{x}_{test}^\top \Omega x_{test} - \hat{x}_{test}^\top \Omega XX^\top (XX^\top + \Sigma)^{-1}XX^\top \Omega x_{test})(\hat{y})

(85)
where the marginal (85) is again given by Bayes’ rule. Assuming zero test noise \( \nu_{\text{test}} = 0 \) and defining the kernel function \( k_{\Omega}(x, x') = x^\top \Omega x' \), we can express the predictive posterior as

\[
\bar{y} \sim \mathcal{N}(k_{\Omega}(x_{\text{test}}, X)(k_{\Omega}(X) + \Sigma)^{-1} y, k_{\Omega}(x_{\text{test}}, X)(k_{\Omega}(X) + \Sigma)^{-1} k_{\Omega}(X, x_{\text{test}}))
\] (86)

### G Lemmas

**Lemma G.1** (Polar coordinates). For any integrable \( f : \mathbb{R} \to \mathbb{R} \),

\[
\int_{\mathbb{R}^2} f(x^2 + y^2) c(x, y) = 2\pi \int_0^\infty f(r^2) r dr
\]

**Proof.** Let \( R = [0, \infty) \times [0, 2\pi] \) and define \( g : \mathbb{R} \to \mathbb{R}^2 \) by \( g(r, \theta) = (r \cos \theta, r \sin \theta) \). Note that \( r^2 = x^2 + y^2 \) and \( g(R) = \mathbb{R}^2 \). The Jacobian of \( g \) at \( (r, \theta) \) is

\[
\begin{vmatrix}
\frac{\partial r \cos \theta}{\partial r} & \frac{\partial r \cos \theta}{\partial \theta} \\
\frac{\partial r \sin \theta}{\partial r} & \frac{\partial r \sin \theta}{\partial \theta}
\end{vmatrix}
= \begin{bmatrix}
\cos \theta & -r \sin \theta \\
\sin \theta & r \cos \theta
\end{bmatrix}
\]

Thus \( |\det(J_g(r, \theta))| = |r (\cos^2 \theta + \sin^2 \theta)| = r \). Thus

\[
\int_{\mathbb{R}^2} f(x^2 + y^2) c(x, y) = \int_R f(g_1(r, \theta)^2 + g_2(r, \theta)^2) |J_g(r, \theta)| d(r, \theta) \tag{by (44)}
\]

\[
= \int_R f(r^2) r d(r, \theta)
\]

\[
= \int_0^\infty \left( \int_0^{2\pi} \exp(-r^2) r d\theta \right) dr \tag{Fubini}
\]

\[
= \int_0^\infty 2\pi \exp(-r^2) r dr 	ag{FTC}
\]

\[
= 2\pi \int_0^\infty \exp(-r^2) r dr \tag{linearity}
\]

\[
\square
\]

**Lemma G.2** (Gaussian integral).

\[
\int_{-\infty}^\infty \exp(-x^2) dx = \sqrt{\pi} \tag{87}
\]

**Proof.** A standard proof shows that \( (\int_{-\infty}^\infty \exp(-x^2) dx)^2 = \pi \) as follows:

\[
\left( \int_{-\infty}^\infty \exp(-x^2) dx \right) \left( \int_{-\infty}^\infty \exp(-y^2) dy \right) = \int_{-\infty}^\infty \int_{-\infty}^\infty \exp(-x^2) dx \exp(-y^2) dy \tag{linearity}
\]

\[
= \int_{-\infty}^\infty \int_{-\infty}^\infty \exp(-x^2) \exp(-y^2) dx dy \tag{linearity}
\]

\[
= \int_{\mathbb{R}^2} \exp(-(x^2 + y^2)) d(x, y) \tag{Fubini}
\]

\[
= 2\pi \int_0^\infty \exp(-r^2) r dr \tag{Lemma G.1}
\]

\[
= 2\pi \left( -\frac{1}{2} \exp(-r^2) \right) \bigg|_0^\infty \tag{FTC}
\]

\[
= 2\pi \left( 0 + \frac{1}{2} \right) = \pi
\]

\[
\square
\]

23
Lemma G.3. For any \( \mu \in \mathbb{R} \) and \( \sigma^2 > 0 \),
\[
\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right) \, dx = 1 \tag{88}
\]

Proof. Let \( u = \frac{x-\mu}{\sqrt{2}\sigma} \) which gives the infinitesimal \( dx = \sqrt{2}\sigma du \). Then
\[
\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right) \, dx = \int_{-\infty}^{\infty} \frac{\sqrt{2\sigma}}{\sqrt{2\pi\sigma}} \exp \left( -u^2 \right) \, du \quad \text{(by (42))}
\]
\[
= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp \left( -u^2 \right) \, du
\]
\[
= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp \left( -u^2 \right) \, du \quad \text{(linearity)}
\]
\[
= 1 \quad \text{(Lemma G.2)}
\]

Lemma G.4.
\[
\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} (x-\mu)^\top \Sigma^{-1} (x-\mu) \right) \, dx = 1
\]

Proof. Let \( \Sigma = U\Lambda U^\top \) denote an orthonormal eigendecomposition. Let \( u = g(x) \) where \( g(x) = \Lambda^{-1/2} U^\top (x - \mu) \). Thus \( |\det(J_g(x))| = |\det(\Lambda^{-1/2} U^\top)| = \det(\Lambda)^{-1/2} \), so we have the infinitesimal \( dx = \sqrt{\det(\Lambda)} du \). Then
\[
\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} (x-\mu)^\top \Sigma^{-1} (x-\mu) \right) \, dx = \int_{\mathbb{R}^d} \frac{\sqrt{\det(\Lambda)}}{(\sqrt{2\pi})^d \sqrt{\det(\Lambda)}} \exp \left( -\frac{1}{2} u^\top u \right) \, du
\]
\[
= \int_{\mathbb{R}^d} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{u_i^2}{2} \right) \, du
\]
By Fubini and linearity,
\[
\int_{\mathbb{R}^d} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{u_i^2}{2} \right) \, du = \int_{-\infty}^{\infty} \left( \cdots \left( \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{u_1^2}{2} \right) \, du_1 \right) \cdots \right) \, du_d
\]
\[
= \prod_{i=1}^{d} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{u_i^2}{2} \right) \, du_i
\]
\[
= \left( \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right) \, dx \right)^d = 1
\]
where the last step applies Lemma G.3 with \( \mu = 0 \) and \( \sigma^2 = 1 \).

Lemma G.5. For any \( \lambda > 0 \), the exponential distribution \( e_\lambda(x) := \lambda \exp(-\lambda x) \) over \([0, \infty)\) has entropy
\[
H(X) = 1 - \log \lambda
\]

Proof.
\[
H(X) = -\int_{0}^{\infty} \lambda \exp(-\lambda x) \log(\lambda \exp(-\lambda x)) \, dx
\]
\[
= -\log \lambda - \lambda \int_{0}^{\infty} \exp(-\lambda x)(-\lambda x) \, dx
\]

24
We evaluate the last integral as follows. Let $u = g(x) = -\lambda x$, then $g'(x) = -\lambda$ so that $|g'(g^{-1}(u))^{-1}| = 1/\lambda$. Reorienting the region between $g(0) = 0$ and $g(\infty) = -\infty$ and applying (43),

\[
\lambda \int_0^\infty \exp(-\lambda x)(-\lambda x)dx = \int_{-\infty}^0 \exp(u)udu = \exp(u)|_0^\infty - \int_{-\infty}^0 \exp(u)du \quad \text{(integration by parts (41))}
\]

\[
= (0 - 0) - \exp(u)|_0^\infty \quad \text{(lim}_{u \to -\infty} \exp(u)u = 0)
\]

\[
= -1
\]

**Lemma G.6.** Define $\Delta := \mu' - \mu$. Then

\[
H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \Delta^\top \Sigma^{-1} \Delta + \frac{1}{2} \text{tr} \left( \Sigma^{-1} \Sigma' \right) + \frac{1}{2} \log((2\pi)^d \det(\Sigma))
\]

*Proof.*

\[
H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) := E_{x \sim \mathcal{N}(\mu', \Sigma')} \left[ \log \mathcal{N}(\mu, \Sigma)(x) \right]
\]

\[
= \frac{1}{2} E_{x \sim \mathcal{N}(\mu', \Sigma')} \left[ (x - \mu)^\top \Sigma^{-1}(x - \mu) \right] + \frac{1}{2} \log((2\pi)^d \det(\Sigma))
\]

By the cyclic property and the linearity of trace,

\[
E_{x \sim \mathcal{N}(\mu', \Sigma')} \left[ (x - \mu)^\top \Sigma^{-1}(X - \mu) \right] = E_{x \sim \mathcal{N}(\mu', \Sigma')} \left[ \text{tr} \left( (x - \mu)^\top \Sigma^{-1}(x - \mu) \right) \right]
\]

\[
= \text{tr} \left( \Sigma^{-1} E_{x \sim \mathcal{N}(\mu', \Sigma')} \left[ (x - \mu)(x - \mu)^\top \right] \right)
\]

Rewriting the expectation,

\[
E_{x \sim \mathcal{N}(\mu', \Sigma')} \left[ (x - \mu)(x - \mu)^\top \right] = E_{x \sim \mathcal{N}(\mu', \Sigma')} \left[ (x - \mu' + \Delta)(x - \mu' + \Delta)^\top \right]
\]

\[
= \text{tr} \left( \Sigma^{-1} \Delta^\top \Sigma^{-1} \Delta + \Delta^\top \Delta \right)
\]

Therefore we have

\[
H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \text{tr} \left( \Sigma^{-1} \Sigma' + \Sigma^{-1} \Delta \Delta^\top \right) + \frac{1}{2} \log((2\pi)^d \det(\Sigma))
\]

\[
= \frac{1}{2} \text{tr} \left( \Sigma^{-1} \Sigma' \right) + \frac{1}{2} \text{tr} \left( \Sigma^{-1} \Delta \Delta^\top \right) + \frac{1}{2} \log((2\pi)^d \det(\Sigma))
\]

**Corollary G.7** (Of Lemma G.6).

\[
H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \log((2\pi)^d \det(\Sigma))
\]

**Corollary G.8** (Of Lemma G.6 and Corollary G.7). Define $\Delta := \mu' - \mu$. Then

\[
\text{KL}(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \Delta^\top \Sigma^{-1} \Delta + \frac{1}{2} \text{tr} \left( \Sigma^{-1} \Sigma' - I_d \times d \right) + \frac{1}{2} \log \frac{\det(\Sigma)}{\det(\Sigma')}
\]
Lemma G.9. Let $A \in \mathbb{R}^{d \times d}$. The main-diagonal block matrix of $A$ at index $k \in \{1 \ldots d\}$ with size $n$ is a matrix $B(k,n) \in \mathbb{R}^{n \times n}$ with entries $B_{i,j}(k,n) = A_{k+i-1,k+j-1}$ for $i,j \in \{1 \ldots n\}$. If $A > 0$, then $B(k,n) > 0$ for all valid $k,n$.

**Proof.** Suppose $u^\top B(k,n)u \leq 0$ for some nonzero $u \in \mathbb{R}^n$. Define $v \in \mathbb{R}^d$ where $v_{k+i-1} = u_i$ for $i = 1 \ldots n$ and other entries are zero. Then $v$ is nonzero and $v^\top Av = u^\top B(k,n)u \leq 0$, contradicting the premise that $A > 0$. \qed

Lemma G.10. Let $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$, $A \in \mathbb{R}^{d \times d}$ and $\Omega \in \mathbb{R}^{d' \times d'}$. Then for all $y \in \mathbb{R}^{d'}$,

$$
E_{X \sim \mathcal{N}(\mu, \Sigma)} \left[ \log \mathcal{N}(AX, \Omega)(y) \right] = \log \mathcal{N}(A\mu, \Omega)(y) - \frac{1}{2} \text{tr} (\Omega^{-1} A \Sigma A^\top)
$$

**Proof.** We have

$$
\log \mathcal{N}(AX, \Omega)(y) = -\frac{1}{2} \log \left( 2\pi \det(\Omega)^{d'} \right) - \frac{1}{2} (y - AX)^\top \Omega^{-1} (y - AX)
$$

where

$$
(y - AX)^\top \Omega^{-1} (y - AX) = \text{tr} \left( (y - AX)^\top \Omega^{-1} (y - AX) \right) = \text{tr} \left( \Omega^{-1} (y - AX) (y - AX)^\top \right) = \text{tr} \left( \Omega^{-1} (yy^\top - yX^\top A^\top - AXy^\top + AXX^\top A^\top) \right)
$$

Since the trace is linear,

$$
E_{X \sim \mathcal{N}(\mu, \Sigma)} \left[ \text{tr} (\Omega^{-1} (yy^\top - yX^\top A^\top - AXy^\top + AXX^\top A^\top)) \right]
= \text{tr} \left( E_{X \sim \mathcal{N}(\mu, \Sigma)} \left[ \Omega^{-1} (yy^\top - yX^\top A^\top - AXy^\top + AXX^\top A^\top) \right] \right)
= \text{tr} \left( \Omega^{-1} (yy^\top - y(A\mu)^\top - A\mu y^\top + AXX^\top A^\top) \right) + \text{tr} (\Omega^{-1} A \Sigma A^\top)
= (y - A\mu)^\top \Omega^{-1} (y - A\mu) + \text{tr} (\Omega^{-1} A \Sigma A^\top)
$$

Thus

$$
E_{X \sim \mathcal{N}(\mu, \Sigma)} \left[ \log \mathcal{N}(AX, \Omega)(y) \right] = -\frac{1}{2} \log \left( 2\pi \det(\Omega)^{d'} \right) - \frac{1}{2} (y - A\mu)^\top \Omega^{-1} (y - A\mu) - \frac{1}{2} \text{tr} (\Omega^{-1} A \Sigma A^\top)
$$

$$
= \log \mathcal{N}(A\mu, \Omega)(y) - \frac{1}{2} \text{tr} (\Omega^{-1} A \Sigma A^\top)
$$

\qed

Lemma G.11. Let $X \sim \mathcal{N}(\mu, \Sigma)$. For any $A \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{m \times d}$,

$$
A\Sigma B^\top = 0_{n \times m} \iff AX \in \mathbb{R}^n \text{ and } BX \in \mathbb{R}^m \text{ are independent}
$$

**Proof.** If $A$ or $B$ is zero then the statement is trivially true (a constant is independent by definition). Otherwise, for all nonzero $(u,v) \in \mathbb{R}^{n+m}$, $(u,v)^\top (AX,BX) = (u^\top A + v^\top B)X$ is normal by the closure under linear transformation (2). Thus $(AX,BX)$ is normal by 4. Hence $AX$ and $BX$ are independent if they are uncorrelated: $E \left[ (AX - \mu)(X - \mu)^\top B^\top \right] = A\Sigma B^\top = 0_{n \times m}$. \qed
Lemma G.12. Let \( X \in \mathbb{R}^d \) and \( Y \in \mathbb{R}^d \) be jointly normal with parameters \((\mu, \Sigma)\). Assume that \( \Sigma_Y - \Sigma_{XY} \Sigma_X^{-1} \Sigma_{XY} \) is invertible. Then for any \( z = (x, y) \in \mathbb{R}^{d+d} \),

\[
\frac{1}{(\sqrt{2\pi})^{d+d} \sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} (z - \mu)^\top \Sigma^{-1} (z - \mu) \right) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma_X)}} \exp \left( -\frac{1}{2} (x - \mu_X)^\top \Sigma_X^{-1} (x - \mu_X) \right) \times \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Omega)}} \exp \left( -\frac{1}{2} (y - \phi(x))^\top \Omega^{-1} (y - \phi(x)) \right) \tag{89}
\]

where \( \Omega \in \mathbb{R}^{d \times d} \) and \( \phi(x) \in \mathbb{R}^d \) are defined as

\[
\Omega := \Sigma_Y - \Sigma_{XY} \Sigma_X^{-1} \Sigma_{XY} \tag{90}
\]

\[
\phi(x) := \mu_Y + \Sigma_{XY} \Sigma_X^{-1} (x - \mu_X) \tag{91}
\]

Proof. By block matrix inversion and abbreviating \( O = \Sigma_X^{-1} \Sigma_{XY} \),

\[
\Sigma^{-1} = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_X^{-1} + O \Omega^{-1} O^\top & -O \Omega^{-1} \\ -\Omega^{-1} O^\top & \Omega^{-1} \end{bmatrix}
\]

Abbreviating \( u = x - \mu_X \) and \( v = y - \mu_Y \),

\[
(z - \mu)^\top \Sigma^{-1} (z - \mu) = u^\top (\Sigma_X^{-1} + O \Omega^{-1} O^\top) u - u^\top O \Omega^{-1} v - v^\top \Omega^{-1} O^\top u + v^\top O^{-1} v
\]

\[
= u^\top \Sigma_X^{-1} u + u^\top O \Omega^{-1} O^\top u - 2 u^\top O \Omega^{-1} v + v^\top \Omega^{-1} v
\]

\[
= u^\top \Sigma_X^{-1} u + (v - O^\top u)^\top \Omega^{-1} (v - O^\top u) = (x - \mu_X)^\top \Sigma_X^{-1} (x - \mu_X) + (y - \phi(x))^\top \Omega^{-1} (y - \phi(x))
\]

where we use the fact that \( \Omega \) is symmetric. By the determinant identity of a block matrix, we have \( \det(\Sigma) = \det(\Sigma_X \Omega) = \det(\Sigma_X) \det(\Omega) \). Applying these identities to the LHS of (89) yields the RHS.

Lemma G.13. Let \((X, Y) \sim \mathcal{N}(\mu, \Sigma)\) where \( X \in \mathbb{R}^d \), \( Y \in \mathbb{R}^d \) and

\[
\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \Lambda_X & \Lambda_{XY} \\ \Lambda_{YX} & \Lambda_Y \end{bmatrix} = \Sigma^{-1}
\]

(i.e., \( \Lambda \) is the precision matrix). Then

\[
Y|X = x \sim \mathcal{N} \left( \mu_Y - \Lambda_Y^{-1} \Lambda_{YX} (x - \mu_X), \Lambda_Y^{-1} \right)
\]

Proof. We can derive the log conditional probability of \( Y = y \) given \( X = x \) from the log joint probability \( X = x, Y = x \) by treating all terms not involving \( y \) as constants. Thus

\[
\log \Pr (Y = y | X = x) = -\frac{1}{2} ((x, y) - \mu)^\top \Sigma^{-1} ((x, y) - \mu) + C
\]

\[
= -\frac{1}{2} ((x, y) - \mu)^\top \Lambda ((x, y) - \mu) + C
\]

\[
= -\frac{1}{2} (y - \mu_Y)^\top \Lambda_Y (y - \mu_Y) - (x - \mu_X)^\top \Lambda_{XY} (y - \mu_Y) + C'
\]

\[
= -\frac{1}{2} y^\top \Lambda_Y y + (\mu_Y^\top \Lambda_Y - (x - \mu_X)^\top \Lambda_{XY}) y + C''
\]

where the key step is directly expanding the precision matrix instead of inverse covariance. By matching the first- and second-order terms in Definition 5, we have \( Y|X = x \sim \mathcal{N}(\nu(x), \Omega) \) where \( \Omega = \Lambda_Y^{-1} \) and \( \nu(x) = \Lambda_Y^{-1} (\Lambda_Y \mu_Y - \Lambda_{YX} (x - \mu_X)) = \mu_Y - \Lambda_{YX} (x - \mu_X) \).
Lemma G.14. Let \( X \sim N(\mu, \Sigma_X) \) and \( Y|X = x \sim N(Ax + b, \Sigma_Y) \) for some \( A \in \mathbb{R}^{d\times d} \) and \( b \in \mathbb{R}^d \). Then

\[
\begin{bmatrix} X \\ Y \end{bmatrix} \sim N \left( \begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_X A^T \\ A\Sigma_X & A\Sigma_X + A\Sigma_X A^T \end{bmatrix} \right)
\]

with the marginal and posterior distributions

\[
Y \sim N(A\mu + b, \Sigma_Y + A\Sigma_X A^T) \quad \quad X|Y = y \sim N(\Lambda_X^{-1}(\Sigma_X^{-1}\mu + A^T\Sigma_Y^{-1}(y - b)), \Lambda_X^{-1}) \quad \quad \Lambda_X^{-1} = (\Sigma_X^{-1} + A^T\Sigma_Y^{-1}A)^{-1}
\]

Proof. With a bit of algebra, we can express the log probability of the variable \( Z = (X, Y) \in \mathbb{R}^{d+d'} \) in the quadratic form:

\[-2 \log \Pr \left( Z = \begin{bmatrix} x \\ y \end{bmatrix} \right) = -2 \log \Pr (X = x) - 2 \log \Pr (Y = y|X = x) \]

\[
= (x - \mu)^T \Sigma_X^{-1}(x - \mu) + (y - Ax - b)^T \Sigma_Y^{-1}(y - Ax - b) + C
\]

\[
= x^T(\Sigma_X^{-1} + A^T\Sigma_Y^{-1}A)x + y^T\Sigma_Y^{-1}y - y^T \Sigma_Y^{-1}Ax - x^T A^T \Sigma_Y^{-1}y
\]

\[+ 2x^T (A^T \Sigma_Y^{-1}b - \Sigma_X^{-1}\mu) - 2y^T \Sigma_Y^{-1}b + C'
\]

\[
= [x^T \quad y^T] \begin{bmatrix} \Sigma_X^{-1} + A^T \Sigma_Y^{-1}A & -A^T \Sigma_Y^{-1} \\ -\Sigma_Y^{-1} A & \Sigma_Y^{-1} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + 2 \begin{bmatrix} A^T \Sigma_Y^{-1}b - \Sigma_X^{-1}\mu \\ -\Sigma_Y^{-1}b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + C'
\]

which shows that the log probability of \( Z = z \) is \(-\frac{1}{2} z^T \Sigma z + \text{constant} \) for some constant \( C'' \in \mathbb{R} \). Thus \( Z \sim N(\nu, \Sigma) \) where \( \Sigma = \Lambda^{-1} \) and \( \nu = \Sigma u \) (Definition 5). The block matrix inversion rule states that

\[
\begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix}^{-1} = \begin{bmatrix} (M_1 - M_2 M_3^{-1} M_4)^{-1} & -(M_1 - M_2 M_3^{-1} M_4)^{-1} M_2 M_4^{-1} \\ -M_4^{-1} M_3 (M_1 - M_2 M_3^{-1} M_4)^{-1} M_2 M_4^{-1} + M_4^{-1} M_3 M_1 M_2 M_4^{-1} & M_4^{-1} - M_4^{-1} M_3 M_1 M_2 M_4^{-1} M_4^{-1} M_2 M_4^{-1} \end{bmatrix}
\]

In this case we have \( M_1 - M_2 M_4^{-1} M_3 = \Sigma_X^{-1} + A^T \Sigma_Y^{-1}A - A^T \Sigma_Y^{-1} \Sigma_Y \Sigma_Y^{-1}A = \Sigma_X^{-1} \) so that

\[
\Sigma = \Lambda^{-1} = \begin{bmatrix} \Sigma_X & \Sigma_X A^T \\ -\Sigma_Y (\Sigma_Y^{-1}A) \Sigma_X & \Sigma_Y + \Sigma_Y (\Sigma_Y^{-1}A) \Sigma_Y (\Sigma_Y^{-1}A) \Sigma_Y \end{bmatrix}
\]

Likewise, the mean is given by a lot of canceling terms:

\[
\nu = \Sigma u = \begin{bmatrix} \Sigma_X & \Sigma_X A^T \\ A \Sigma_X & \Sigma_Y + A \Sigma_X A^T \end{bmatrix} \begin{bmatrix} \Sigma_X^{-1} \mu - A^T \Sigma_Y^{-1}b \\ \Sigma_Y^{-1}b \end{bmatrix} = \begin{bmatrix} \mu \\ A \mu + b \end{bmatrix}
\]

This shows the statement about the marginal probability of \( Y \). To obtain the statement about the posterior probability of \( X \) given \( Y = y \), we use the precision matrix form which states that (swapping \( X \) and \( Y \) in (8))

\[
X|Y = x \sim N(\mu_X - \Lambda_X^{-1} \Lambda_{XY}(y - \mu_Y), \Lambda_X^{-1})
\]

where \( \mu_X = \mu, \mu_Y = A\mu + b, \Lambda_X = \Sigma_X^{-1} + A^T \Sigma_Y^{-1}A, \) and \( \Lambda_{XY} = -A^T \Sigma_Y^{-1} \). Noting \( \Lambda_X = \Sigma_X^{-1} - \Lambda_{XY} \), we can simplify the mean as

\[
\mu - \Lambda_X^{-1} \Lambda_{XY}(y - A\mu - b) = \Lambda_X^{-1}(\Lambda_X \mu - \Lambda_{XY}(y - A\mu - b)) = \Lambda_X^{-1}(\Sigma_X^{-1}\mu + A^T \Sigma_Y^{-1}(y - b))
\]

Lemma G.15. Let \( X \in \mathbb{R}^d \) and \( Y \in \mathbb{R}^{d'} \) be jointly normal with parameters \((\mu, \Sigma)\). Assume that \( \Sigma_Y - \Sigma_Y \Sigma_X^{-1} \Sigma_X Y \) is invertible. Then for any \( x \in \mathbb{R}^d \),

\[
H(Y|X = x) = \frac{1}{2} \log \left( (2\pi e)^d \det(\Sigma_Y - \Sigma_Y \Sigma_X^{-1} \Sigma_X Y) \right)
\]

\[
I(X,Y) = \frac{1}{2} \log \left( \frac{\det(\Sigma_X) \det(\Sigma_Y)}{\det(\Sigma)} \right)
\]
Proof. By Lemma G.12, \( Y|X = x \) is distributed as \( \mathcal{N}(\phi(x), \Omega) \) for any \( x \in \mathbb{R}^d \) where \( \phi(x) := \mu_Y + \Sigma_{YX} X^{-1}(x - \mu_X) \) and \( \Omega := \Sigma_Y - \Sigma_{YX} X^{-1} \Sigma_{XY} \). Thus

\[
H(Y|X = x) = \mathbb{E}[-\log \Pr(Y|X = x)] = \frac{1}{2} \mathbb{E}[(Y - \phi(x))^\top \Omega^{-1}(Y - \phi(x))] + \frac{1}{2} \log((2\pi)^d \det(\Omega))
\]

Using the cyclic property and linearity of trace, the first term is

\[
\frac{1}{2} \mathbb{E}[(Y - \phi(x))^\top \Omega^{-1}(Y - \phi(x))] = \frac{1}{2} \text{tr} \left( \Omega^{-1} \mathbb{E}[(Y - \phi(x))(Y - \phi(x))^\top] \right) = \frac{d'}{2}
\]

This shows (92) (note the Euler constant \( e \)). To show (93), we have

\[
I(X, Y) = H(Y) - H(Y|X)
\]

\[
= \frac{1}{2} \log \left( (2\pi)^d \det(\Sigma_Y) \right) - \frac{1}{2} \log \left( (2\pi)^d \det(\Omega) \right)
\]

\[
= \frac{1}{2} \log \left( \frac{\det(\Sigma_Y)}{\det(\Omega)} \right)
\]

where for the last equality we use the fact that \( \det(\Sigma) = \det(\Sigma_Y) = \det(\Sigma_X \Omega) = \det(\Sigma_X) \det(\Omega) \).

\[\square\]

**Lemma G.16.** The following statements about \( X \in \mathbb{R}^d \) are equivalent.

1. \( X \sim \mathcal{N}(\mu, \Sigma) \), that is, \( \Pr(X = x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)) \).
2. \( M_X(t) = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t) \) for all \( t \in \mathbb{R}^d \).
3. \( X = \Sigma^{1/2} Z + \mu \) where \( Z \sim \mathcal{N}(0_d, I_{d \times d}) \).
4. \( Y = a^\top X \) has the density \( \mathcal{N}(a^\top \mu, a^\top \Sigma a) \) for all nonzero \( a \in \mathbb{R}^d \).
5. \( \log \Pr(X = x) = -\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) + C = -\frac{1}{2} x^\top \Sigma^{-1} x + (\Sigma^{-1} \mu)^\top x + C' \) for some constants \( C, C' \in \mathbb{R} \).

**Proof.** Lemma C.1 gives 1 ⇔ 2. To show 2 ⇔ 3 we note that by (65)

\[
M_{\Sigma^{1/2} Z + \mu}(t) = \exp(t^\top \mu) M_Z(\Sigma^{1/2} t) = \exp \left( t^\top \mu + \frac{1}{2} t^\top \Sigma t \right) = M_X(t)
\]

We have 1 ⇒ 4 since the density of \( Y \) is \( \mathcal{N}(a^\top \mu, a^\top \Sigma a) \) by Lemma C.2. To show 4 ⇒ 2, pick any nonzero \( a \in \mathbb{R}^d \). For all \( t \in \mathbb{R} \)

\[
M_X(ta) = M_{a^\top X}(t) = \exp \left( ta^\top \mu + \frac{1}{2} t^2 a^\top \Sigma a \right)
\]

where the first equality uses (65) and the second equality uses Lemma C.1. Setting \( t = 1 \) gives \( M_X(a) = \exp(a^\top \mu + \frac{1}{2} a^\top \Sigma a) \). Additionally, \( M_X(0_d) = 1 = \exp(0_d^\top \mu + \frac{1}{2} 0_d^\top \Sigma 0_d) \). Thus \( M_X(t) = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t) \) for all \( t \in \mathbb{R}^d \). To show 1 ⇒ 5, the log probability of \( X = x \) where \( X \sim \mathcal{N}(\mu, \Sigma) \) is

\[
\log \Pr(X = x) = -\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) + C_1
\]

\[
= -\frac{1}{2} x^\top \Sigma^{-1} x + \mu^\top \Sigma^{-1} x - \frac{1}{2} \mu^\top \Sigma^{-1} \mu + C_1
\]

("expanding the square")

\[
= -\frac{1}{2} x^\top \Sigma^{-1} x + (\Sigma^{-1} \mu)^\top x + C_2
\]

where \( C_1, C_2 \in \mathbb{R} \) are some constants independent of \( x \). To show 5 ⇒ 1, note that

\[
\log \Pr(X = x) = -\frac{1}{2} x^\top \Sigma^{-1} x + (\Sigma^{-1} \mu)^\top x + C
\]

\[
= -\frac{1}{2} x^\top \Sigma^{-1} x + \mu^\top \Sigma^{-1} x - \frac{1}{2} \mu^\top \Sigma^{-1} \mu + \frac{1}{2} \mu^\top \Sigma^{-1} \mu + C
\]

\[
= -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) + C'
\]

("completing the square")

29
where \( C' \) is a constant. This implies, for the constant \( C'' = \exp(C') \),

\[
\Pr(X = x) = C'' \exp \left( -\frac{1}{2} (x - \mu) \Sigma^{-1} (x - \mu) \right)
\]

Since \( \int_{x \in \mathbb{R}^d} \Pr(X = x) \, dx = 1 \), we have

\[
C'' = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}
\]

where the second equality is by Lemma G.4.

**Lemma G.17** (Popoviciu’s inequality). For any bounded scalar random variable \( X \in [a, b] \),

\[
\text{Var}(X) \leq \frac{(b - a)^2}{4}
\]

with equality iff \( \Pr(X = a) = \Pr(X = b) = \frac{1}{2} \).

**Proof.** For any constant \( c \in \mathbb{R} \),

\[
\mathbb{E}[(X - c)^2] = \mathbb{E}[(X - \mathbb{E}[X] + \mathbb{E}[X] - c)^2] \geq \text{Var}(X).
\]

Choosing \( c = \frac{b + a}{2} \) and using the fact that \( |X - \frac{b + a}{2}| \leq \frac{b - a}{2} \), we have \( \text{Var}(X) \leq \mathbb{E}[(X - \frac{b + a}{2})^2] \leq \frac{(b - a)^2}{4} \). \( \square \)

**Lemma G.18** (Markov’s inequality). For any nonnegative scalar random variable \( X \geq 0 \), for any \( \epsilon > 0 \):

\[
\Pr(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}
\]

**Proof.**

\[
\mathbb{E}[X] = \int_0^\infty \Pr(X = x) \, dx \quad \text{(proof similar if } X \text{ is discrete)}
\]

\[
\geq \int_\epsilon^\infty \Pr(X = x) \, dx
\]

\[
\geq \int_\epsilon^\infty \Pr(X = x) \, dx \epsilon \, dx
\]

\[
\geq \epsilon \Pr(X \geq \epsilon)
\]

\( \square \)

**Lemma G.19** (Chernoff’s inequality). For any scalar random variable \( X \in \mathbb{R} \) and \( \epsilon \geq \mathbb{E}[X] \),

\[
\Pr(X \geq \epsilon) \leq e^{-\psi_X^*(\epsilon)}
\]

where \( \psi_X^*(\epsilon) = \sup_{t \in \mathbb{R}} t \epsilon - \psi_X(t) \) is the Legendre transform of the CGF \( \psi_X(t) = \log \mathbb{E}[e^{tX}] \).

**Proof.**

\[
\Pr(X \geq \epsilon) \leq \Pr(tX \geq t\epsilon)
\]

\[
= \Pr(e^{tX} \geq e^{t\epsilon})
\]

\[
= \frac{\mathbb{E}[e^{tX}]}{e^{t\epsilon}} \quad \text{(Markov’s inequality, since } e^{tX} \geq 0 \text{ and } e^{t\epsilon} > 0)
\]

In particular,

\[
\Pr(X \geq \epsilon) \leq \inf_{t \geq 0} e^{-(t\epsilon - \psi_X(t))}
\]

\[
= e^{-(\sup_{t \geq 0} t\epsilon - \psi_X(t))}
\]

\[
= e^{-(\sup_{t \in \mathbb{R}} t\epsilon - \psi_X(t))} \quad \text{(94)}
\]

The step (94) uses the following lemma.
Lemma. Let $J(t) := \psi'(t) - \psi(t)$ and $J^* = \sup_{t \in \mathbb{R}} J(t)$. Then $J^* \geq J(0)$.

Proof.

\begin{align*}
J(t) &= te - \log \mathbf{E}[e^{tX}] \\
&\leq te - t\mathbf{E}[X] \\
&= t(\epsilon - \mathbf{E}[X]) \quad (\text{Jensen’s inequality: } \log \mathbf{E}[X] \geq \mathbf{E}[\log X] \geq 0)
\end{align*}

Thus $J(t) \leq 0$ for all $t < 0$. The lemma follows from the fact that $J(0) = 0$.

\[\square\]

**Theorem G.20** (Factorization Theorem). Assume a joint distribution

\[p_{\Theta X T}(\theta, x, t) = p_{\Theta}(\theta) \times p_{X|\Theta}(x|\theta) \times [\tau(x) = t]\]

where $X \in \mathcal{X}$ is a sample from a distribution parameterized by $\Theta \in \mathcal{H}$, and $T = \tau(X) \in T$ is the sample statistic for some function $\tau : \mathcal{X} \to T$. The following statements about $\tau$ are equivalent: if any holds, we say $\tau$ is a **sufficient statistic** for $\Theta$.

- $X$ is conditionally independent of $\Theta$ given $T = t$:
  \[p_{X|T}(x|t) = p_{X|T \Theta}(x|t, \theta)\]  \hspace{1cm} (95)

- There exist $f_T : \mathcal{X} \to \mathbb{R}$ and $g : T \times \mathcal{H} \to \mathbb{R}$ such that
  \[p_{X|\Theta}(x|\theta) = f_T(x) \times g(\tau(x), \theta)\]  \hspace{1cm} (96)

Proof. (96)⇒(95): For any $t, \theta$,

\[p_{T|\Theta}(t|\theta) = \sum_{x \in \mathcal{X} : \tau(x) = t} p_{X|\Theta}(x|\theta) \quad (\text{proof similar if } X \text{ is continuous})
\]

\[= \sum_{x \in \mathcal{X} : \tau(x) = t} f_T(x) \times g(\tau(x), \theta) \quad (96)
\]

\[= \left( \sum_{x \in \mathcal{X} : \tau(x) = t} f_T(x) \right) \times g(t, \theta)
\]

thus for any $x$ satisfying $\tau(x) = t$,

\[p_{X|T \Theta}(x|t, \theta) = \frac{p_{X T \Theta}(x, t|\theta)}{p_{T|\Theta}(t|\theta)} = \frac{f_T(x) \times g(t, \theta)}{\sum_{x \in \mathcal{X} : \tau(x) = t} f_T(x)} \times g(t, \theta) = \frac{f_T(x)}{\sum_{x \in \mathcal{X} : \tau(x) = t} f_T(x)}
\]

and $p_{X|T \Theta}(x|t, \theta) = 0$ for $x$ such that $\tau(x) \neq t$. This implies $p_{X|T}(x|t) = p_{X|T \Theta}(x|t, \theta)$ for all $\theta$.

(95)⇒(96): Define $f_T(x) = p_{X|T}(x|\tau(x))$ and $g(t, \theta) = p_{T|\Theta}(t|\theta)$. Then

\[p_{X|\Theta}(x|\theta) = p_{X T \Theta}(x, \tau(x)|\theta) = p_{X|T}(x|\tau(x)) \times p_{T|\Theta}(t|\theta)
\]

\[= p_{X|T}(x|\tau(x)) \times p_{T|\Theta}(\tau(x)|\theta)
\]

\[= f_T(x) \times g(\tau(x), \theta) \quad (95)
\]

\[\square\]

**Lemma G.21.** Let $X \in \mathcal{X}$ be a random variable and $\tau : \mathcal{X} \to \mathbb{R}^m$ be a function such that

\[B_{p,\tau}(\theta) := \log \mathbf{E} \left[ e^{\theta^T \tau(X)} \right]
\]

exists for all $\theta \in \mathbb{R}^m$. Then $B_{p,\tau} : \mathbb{R}^m \to \mathbb{R}$ is convex.
Proof. We use Hölder’s inequality which states that $\mathbf{E}[|XY|] \leq \mathbf{E}[|X|^p]^{\frac{1}{p}} \mathbf{E}[|Y|^q]^{\frac{1}{q}}$ for any $p, q \geq 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$. For any $\alpha \in [0,1]$ and $\theta, \omega \in \mathbb{R}^m$:

$$
\exp (B_{p,\tau}(\alpha \theta + (1 - \alpha)\omega)) = \mathbf{E} \left[ e^{\alpha \theta^\top \tau(X) + (1 - \alpha)\omega^\top \tau(X)} \right] \\
= \mathbf{E} \left[ e^{\alpha \theta^\top \tau(X)} \left| e^{(1 - \alpha)\omega^\top \tau(X)} \right|^\frac{1}{1 - \alpha} \right]^{(1 - \alpha)} \\
\leq \mathbf{E} \left[ e^{\alpha \theta^\top \tau(X)} \right]^\alpha \mathbf{E} \left[ e^{(1 - \alpha)\omega^\top \tau(X)} \right]^{\frac{1}{1 - \alpha}}^{(1 - \alpha)} \\
= \exp (B_{p,\tau}(\theta))^\alpha \exp (B_{p,\tau}(\omega))^{(1 - \alpha)}
$$

Taking the log on both sides yields $B_{p,\tau}(\alpha \theta + (1 - \alpha)\omega) \leq \alpha B_{p,\tau}(\theta) + (1 - \alpha)B_{p,\tau}(\omega)$.

\[\square\]

Lemma G.22. Let $p$ be a distribution over $\mathbb{R}^d$ and define $q_{p,\tau,\theta}(x) := \frac{e^{\theta^\top \tau(x)p(x)}}{\mathbf{E}_{x' \sim p}[e^{\theta^\top \tau(x')}]}$ for function $\tau : \mathbb{R}^d \rightarrow \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ where $\mathbf{E}_{x' \sim p}[e^{\theta^\top \tau(x')}]$ exists. Let $B_{p,\tau}(\theta) := \log \mathbf{E}_{x \sim p}[e^{\theta^\top \tau(x)}]$. Then

$$
\nabla B_{p,\tau}(\theta) = \mathbf{E}_{x \sim q_{p,\tau,\theta}} [\tau(x)] \\
\nabla^2 B_{p,\tau}(\theta) = \mathbf{Cov} (\tau(x))
$$

Proof.

$$
\nabla B_{p,\tau}(\theta) = \frac{\mathbf{E}_{x \sim p}[e^{\theta^\top \tau(x)} \tau(x)]}{\mathbf{E}_{x \sim p}[e^{\theta^\top \tau(x)}]} \\
\nabla^2 B_{p,\tau}(\theta) = \frac{\mathbf{E}_{x \sim p}[e^{\theta^\top \tau(x)} \tau(x) \tau(x)^\top]}{\mathbf{E}_{x \sim p}[e^{\theta^\top \tau(x)}]} - \left( \frac{\mathbf{E}_{x \sim p}[e^{\theta^\top \tau(x)} \tau(x)]}{\mathbf{E}_{x \sim p}[e^{\theta^\top \tau(x)}]} \right)^\top
$$

Thus by the definition of $q_{p,\tau,\theta}$

$$
\nabla B_{p,\tau}(\theta) = \mathbf{E}_{x \sim q_{p,\tau,\theta}} [\tau(x)] \\
\nabla^2 B_{p,\tau}(\theta) = \mathbf{E}_{x \sim q_{p,\tau,\theta}} [\tau(x) \tau(x)^\top] - \left( \mathbf{E}_{x \sim q_{p,\tau,\theta}} [\tau(x)] \right)^\top
$$

\[\square\]

Lemma G.23. $\mathcal{N}(\mu, \Sigma)$ is in the exponential family, with one parameterization given by

$$
h(x) = \frac{1}{(\sqrt{2\pi})^d} \\
\theta = \begin{bmatrix} \Sigma^{-1}\mu \\ \frac{1}{2} \text{vec} (\Sigma^{-1}) \end{bmatrix} \in \mathbb{R}^{d(d + 1)} \\
\tau(x) = \begin{bmatrix} x \\ \text{vec} \left( xx^\top \right) \end{bmatrix} \in \mathbb{R}^{d(d + 1)} \\
A_{h,\tau}(\theta) = \frac{1}{2} \left( \mu^\top \Sigma^{-1}\mu + \log (\text{det}(\Sigma)) \right)
$$

where $\text{vec}(M) \in \mathbb{R}^{n^2}$ is the vector form of matrix $M \in \mathbb{R}^{n \times n}$ with $\text{vec}(M)_{(i-1)n+j} = M_{i,j}$.
Proof.

\[ \mathcal{N}(\mu, \Sigma)(x) = \frac{1}{(\sqrt{2\pi})^d\sqrt{|\det(\Sigma)|}} \exp \left( -\frac{1}{2} (x - \mu)\Sigma^{-1}(x - \mu) \right) \]

\[ = \frac{1}{(\sqrt{2\pi})^d} \exp \left( \mu^\top \Sigma^{-1}x - \frac{1}{2} x^\top \Sigma^{-1}x - \frac{1}{2} \mu^\top \Sigma^{-1}\mu - \frac{1}{2} \log(\det(\Sigma)) \right) \]

\[ = \frac{1}{(\sqrt{2\pi})^d} \exp \left( \left[ -\frac{1}{2} \text{vec}(\Sigma^{-1}) \right]^\top \left[ \begin{bmatrix} x \\ \text{vec}(x^\top) \end{bmatrix} \right] - \frac{1}{2} \left( \mu^\top \Sigma^{-1}\mu + \log(\det(\Sigma)) \right) \right) \]

where we use the fact that \( u^\top M v = \text{vec}(M)^\top \text{vec}(uw^\top) \). \( \square \)

**Lemma G.24.** Let \( q_{h,\tau,\theta}(x) = h(x) \exp(\theta^\top \tau(x) - A_{h,\tau}(\theta)) \) with \( A_{h,\tau}(\theta) = \log \left( \int_{x \in \mathbb{R}^d} h(x) \exp(\theta^\top \tau(x)) \, dx \right) \) denote an exponential family. The log-MGF of the sufficient statistic \( \tau(x) \) is given by

\[ \psi_{\tau(x)}(t) = A_{h,\tau}(\theta + t) - A_{h,\tau}(\theta) \]

**Proof.**

\[ M_{\tau(x)}(t) = \mathbb{E}_{x \sim q_{h,\tau,\theta}} \left[ \exp(t^\top \tau(x)) \right] \]

\[ = \int_{x \in \mathbb{R}^d} h(x) \exp(\theta^\top \tau(x) - A_{h,\tau}(\theta)) \exp(t^\top \tau(x)) \, dx \]

\[ = \exp(-A_{h,\tau}(\theta)) \int_{x \in \mathbb{R}^d} h(x) \exp(\theta^\top (\tau(x) + t)) \, dx \]

\[ = \exp(A_{h,\tau}(\theta + t) - A_{h,\tau}(\theta)) \]

**Lemma G.25.** Let \( q_{h,\tau,\theta}(x) = h(x) \exp(\theta^\top \tau(x) - A_{h,\tau}(\theta)) \) with \( A_{h,\tau}(\theta) = \log \left( \int_{x \in \mathbb{R}^d} h(x) \exp(\theta^\top \tau(x)) \, dx \right) \) denote an exponential family. Define a distribution over \( \theta \in \mathbb{R}^m \) by

\[ \pi_{h,\tau}(\theta; \alpha, \beta) := \frac{1}{Z_{h,\tau}(\alpha, \beta)} \exp(\theta^\top \alpha - \beta A_{h,\tau}(\theta)) \]

for \( \alpha \in \mathbb{R}^m \) and \( \beta \in \mathbb{R} \) such that \( Z_{h,\tau}(\alpha, \beta) := \int_{\theta \in \mathbb{R}^m} \exp(\theta^\top \alpha - \beta A_{h,\tau}(\theta)) \, d\theta \) exists. Then the conditional distribution over \( \theta \) given \( x \) is

\[ \kappa_{h,\tau}(\theta|x; \alpha, \beta) = \pi_{h,\tau}(\theta; \tau(x) + \alpha, 1 + \beta) \]

**Proof.** By Bayes' rule,

\[ \kappa_{h,\tau}(\theta|x; \alpha, \beta) \propto \pi_{h,\tau}(\theta; \alpha, \beta) \times q_{h,\tau,\theta}(x) \]

\[ = \frac{1}{Z_{h,\tau}(\alpha, \beta)} \exp(\theta^\top \alpha - \beta A_{h,\tau}(\theta)) \times h(x) \exp(\theta^\top \tau(x) - A_{h,\tau}(\theta)) \]

\[ \times \exp(\theta^\top (\tau(x) + \alpha) - (1 + \beta)A_{h,\tau}(\theta)) \]

This implies \( \kappa_{h,\tau}(\theta|x; \alpha, \beta) = \pi_{h,\tau}(\theta; \tau(x) + \alpha, 1 + \beta) \). \( \square \)

**Lemma G.26.** Let \( X_t \) denote the \( t \)-tilted \( X \sim \mathcal{N}(\mu, \Sigma) \) using \( \tau(x) = x \). Then

\[ X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma) \]
Proof. We can directly verify this claim using the fact that the CGF of $X$ is $\mu^T t + \frac{1}{2} t^T \Sigma t$:

$$\Pr(X_t = x) = \frac{e^{t^T x}}{\mathbb{E}[e^{t^T X}]} = \mathcal{N}(\mu, \Sigma)(x)$$

where $\mathcal{N}(\mu, \Sigma)$ is the Gaussian distribution.

$$= \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} (x - \mu) \Sigma^{-1} (x - \mu) + t^T x - \mu^T t - \frac{1}{2} t^T \Sigma t \right)$$

$$= \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} (x - \mu - \Sigma t) \Sigma^{-1} (x - \mu - \Sigma t) \right)$$

\[\square\]

**Lemma G.27** (Hoeffding’s lemma). Let $X \in [a, b]$ be a bounded scalar random variable. Then

$$\psi_{X-\mathbb{E}[X]}(t) \leq \frac{(b-a)^2 t^2}{8}$$

**Proof.** For any $t \in \mathbb{R}$, by Taylor’s approximation of $\psi_X$ around 0, for some $\eta$ between 0 and $t$:

$$\psi_X(t) = \psi_X(0) + \frac{1}{2} \frac{\psi''(0)}{\mathbb{E}[X^2]} t^2 \quad \iff \quad \psi_{X-\mathbb{E}[X]}(t) = \frac{\text{Var}(X)}{2}$$

where $X_\eta \in [a, b]$ is the $\eta$-tilted $X$ (67). By Popoviciu’s inequality (Lemma G.17), $\text{Var}(X_\eta) \leq \frac{(b-a)^2}{4}$.

\[\square\]

**Lemma G.28.** Let $\psi^*_X(t) := \sup_{\lambda \in \mathbb{R}^d} \lambda^T t - \psi_X(\lambda)$ denote the Legendre transform of $\psi_X$. If $X \sim \mathcal{N}(\mu, \Sigma)$,

$$\psi^*_X(t) = \frac{1}{2} (t - \mu)^T \Sigma^{-1} (t - \mu)$$

**Proof.** $J(\lambda) = \lambda^T t - \psi_X(\lambda)$ is concave in $\lambda \in \mathbb{R}^d$ since $\psi_X$ is convex. The stationary condition is

$$\nabla J(\lambda) = t - \nabla \psi_X(\lambda) = t - \mu - \Sigma \lambda = 0_d$$

Thus $\lambda^\ast = \Sigma^{-1}(t - \mu)$ is the maximizer of $J$. Then

$$\psi^*_X(t) = (\lambda^\ast)^T t - \psi_X(\lambda^\ast)$$

$$= (\lambda^\ast)^T t - (\lambda^\ast)^T \mu - \frac{1}{2} (\lambda^\ast)^T \Sigma \lambda^\ast$$

$$= (t - \mu)^T \Sigma^{-1} t - (t - \mu)^T \Sigma^{-1} \mu - \frac{1}{2} (t - \mu)^T \Sigma^{-1} (t - \mu)$$

$$= \frac{1}{2} (t - \mu)^T \Sigma^{-1} (t - \mu)$$

\[\square\]

**Lemma G.29.** If $X \sim \mathcal{G}(\sigma^2)$, then $\text{Var}(X) \leq \sigma^2$.

**Proof.** By the Taylor series of $e^z = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots$,

$$f(t) := \mathbb{E}[e^{tX}] = \mathbb{E} \left[ 1 + tX + \frac{t^2 X^2}{2} + \frac{t^3 X^3}{6} + \cdots \right] = 1 + \frac{t^2 \mathbb{E}[X^2]}{2} + t^3 P_1(t)$$

$$g(t) := \mathbb{E}[e^{\frac{a^2 t^2}{2}}] = 1 + \frac{a^2 t^2}{2} + \frac{a^4 t^4}{4} + \cdots = 1 + \frac{a^2 t^2}{2} + t^3 P_2(t)$$

where $P_1, P_2$ are some polynomials. By premise, for all $t \in \mathbb{R}$

$$f(t) \leq g(t) \quad \iff \quad \frac{t^2 \mathbb{E}[X^2]}{2} + t^3 P_1(t) \leq \frac{a^2 t^2}{2} + t^3 P_2(t)$$

\[\Rightarrow \quad \mathbb{E}[X^2] - \sigma^2 \leq t G(t)\]
where $G$ is again some polynomial. Thus

$$\mathbf{E}[X^2] - \sigma^2 \leq \lim_{t \to 0} tG(t) = 0 \quad \iff \quad \mathbf{E}[X^2] \leq \sigma^2$$

\[ \square \]

**Lemma G.30.** If $X, Z \in \mathbb{R}$ are random variables with the CGFs $\psi_X, \phi_Z : \mathbb{R} \to \mathbb{R}$,

$$\psi_X(t) \leq \phi_Z(t) \quad \forall t \in \mathbb{R} \quad \Rightarrow \quad \exp(-\psi_X(t)) \leq \exp(-\phi_Z(t)) \quad \forall t \in \mathbb{R}$$

where $\psi_X^*(t) = \sup_{\lambda \in \mathbb{R}} \lambda t - \psi_X(t)$ is the Legendre transform of $\psi_X$ (similarly for $\phi_Z^*(t)$).

**Proof.**

$$\psi_X(t) \leq \phi_Z(t) \quad \iff \quad -\psi_X(t) \geq -\phi_Z(t)$$

$$\iff \lambda t - \psi_X(t) \geq \lambda t - \phi_Z(t) \quad \forall \lambda \in \mathbb{R}$$

$$\iff \sup_{\lambda \in \mathbb{R}} \lambda t - \psi_X(t) \geq \sup_{\lambda \in \mathbb{R}} \lambda t - \phi_Z(t)$$

$$\iff \psi_X^*(t) \geq \phi_Z^*(t)$$

$$\iff -\psi_X^*(t) \leq -\phi_Z^*(t)$$

$$\iff \exp(-\psi_X^*(t)) \leq \exp(-\phi_Z^*(t))$$

\[ \square \]

**Lemma G.31.** If $X_1 \ldots X_N$ are independently sub-Gaussian with $X_i \sim \mathcal{G}(\sigma_i^2)$, then for all $\epsilon \geq 0$:

$$\Pr \left( \left| \frac{1}{N} \sum_{i=1}^{N} X_i \right| \geq \epsilon \right) \leq 2 \exp \left( -\frac{N^2 \epsilon^2}{2 \left( \sum_{i=1}^{N} \sigma_i^2 \right)} \right)$$

**Proof.** Let $\tilde{X} := \sum_{i=1}^{N} X_i$. Note that $\tilde{X} \sim \mathcal{G}(\sum_{i=1}^{N} \sigma_i^2)$ (4) and $-\tilde{X} \sim \mathcal{G}(\sum_{i=1}^{N} \sigma_i^2)$ (2). Thus

$$\Pr \left( \left| \frac{1}{N} \tilde{X} \right| \geq \epsilon \right) = \Pr \left( \frac{1}{N} \tilde{X} \leq -\epsilon \lor \frac{1}{N} \tilde{X} \geq \epsilon \right)$$

$$\leq \Pr \left( \frac{1}{N} \tilde{X} \leq -\epsilon \right) + \Pr \left( \frac{1}{N} \tilde{X} \geq \epsilon \right) \quad \text{(union bound)}$$

$$= \Pr \left( -\tilde{X} \geq N\epsilon \right) + \Pr \left( \tilde{X} \geq N\epsilon \right)$$

$$\leq 2 \exp \left( -\frac{N^2 \epsilon^2}{2 \left( \sum_{i=1}^{N} \sigma_i^2 \right)} \right) \quad \text{(3)}$$

\[ \square \]

**Lemma G.32.** The gradient $\nabla \mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to \mathbb{R}^d$ and the Hessian $\nabla^2 \mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ of $\mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to [0, 1]$ are

$$\nabla \mathcal{N}(\mu, \Sigma)(x) = -\mathcal{N}(\mu, \Sigma)(x) \times \Sigma^{-1}(x - \mu)$$

$$\nabla^2 \mathcal{N}(\mu, \Sigma)(x) = -\mathcal{N}(\mu, \Sigma)(x) \times (\Sigma^{-1} - \Sigma^{-1}(x - \mu)(x - \mu)^\top \Sigma^{-1})$$

The Hessian is negative-definite at $x = \mu$, but possibly indefinite at other points.

**Proof.** Shorthanding $p(x) = \mathcal{N}(\mu, \Sigma)(x)$, $C = ((\sqrt{2\pi})^d \sqrt{\det(\Sigma)})^{-1}$, and $g(x) = -\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)$ where $\nabla g(x) = -\Sigma^{-1}(x - \mu)$, we have

$$\nabla p(x) = C \nabla \exp(g(x)) = C \exp(g(x)) \nabla g(x) = p(x) \left( -\Sigma^{-1}(x - \mu) \right) = p(x)k(x)$$

35
with \(k(x) = -\Sigma^{-1}(x - \mu)\). Denoting the Jacobian \(J_k(x) = -\Sigma^{-1}\), we have

\[
\nabla^2 p(x) = p(x)J_k(x) + k(x)(\nabla p(x)) = -p(x)\Sigma^{-1} + p(x)k(x)k(x)^\top = -p(x)\left(\Sigma^{-1} - \Sigma^{-1}(x - \mu)(x - \mu)^\top\Sigma^{-1}\right)
\]

Now we analyze the Hessian \(H(x) = \nabla^2 p(x)\). To show that it is negative-definite at \(x = \mu\), we simply note that

\[
H(\mu) = -p(\mu)\Sigma^{-1} < 0 \quad \forall \lambda > 0, \forall \mu > 0
\]

(the inverse of a positive-definite matrix \(\Sigma\) remains positive-definite). For the last statement, it is sufficient to give an example of an indefinite Hessian. Let \(\mu = (0, 0)\) and \(\Sigma = I_{2 \times 2}\) (i.e., standard Gaussian in \(d = 2\) dimensions). Pick the point \(x = (1, 1)\), one standard deviation away from the mean in each dimension. For any vector \(u \in \mathbb{R}^2\), we have

\[
u^\top H(x)u = -p(x)\left[u^\top \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} u\right]
\]

where the matrix is indefinite. For instance, \(u = (1, -1)\) results in \(-2p(x) < 0\) and \(u = (1, 1)\) results in \(2p(x) > 0\).

In the two-dimensional case, it is easy to visualize that this is a saddle point (convex along the direction of \((1, 1)\), concave along the orthogonal direction). Interestingly, for \(d = 1\), the points \(x = \mu \pm \sigma\) result in \(H(x) = 0\) ("inflection points").

Lemma G.33. \(\frac{\partial}{\partial a} \Phi(\sqrt{a})\Big|_{a=0} = \text{sigmoid}'(0) = \frac{1}{4}\)

Proof. We have \(\text{sigmoid}'(a) = \text{sigmoid}(a)(1 - \text{sigmoid}(a))\) and \(\text{sigmoid}(0) = \frac{1}{2}\), so \(\text{sigmoid}'(0) = \frac{1}{4}\). On the other hand, we have

\[
\frac{\partial \Phi(\lambda a)}{\partial a} = \lambda N(0, 1)(\lambda a) \quad \Rightarrow \quad \frac{\partial \Phi(\lambda a)}{\partial a} \bigg|_{a=0} = \lambda N(0, 1)(0) = \frac{\lambda}{\sqrt{2\pi}}
\]

Matching the two values yield \(\lambda = \sqrt{\frac{1}{2}}\).

Lemma G.34. For any \(\lambda, \beta \in \mathbb{R}\),

\[
\mathbb{E}_{X \sim N(\mu, \sigma^2)}[\Phi(\lambda X + \beta)] = \Phi\left(\frac{\lambda \mu + \beta}{\sqrt{1 + \lambda^2 \sigma^2}}\right)
\]

Proof.

\[
\mathbb{E}_{X \sim N(\mu, \sigma^2)}[\Phi(\lambda X + \beta)] = \mathbb{E}_{Z \sim N(0, 1)}[\Phi(\lambda \mu + \lambda \sigma Z + \beta)] \\
= \Pr_{Z, Z' \sim N(0, 1)}(Z' < \lambda \mu + \lambda \sigma Z + \beta) \\
= \Pr_{Z, Z' \sim N(0, 1)}(Z' - \lambda \sigma Z < \lambda \mu + \beta) \\
= \Pr_{Z'' \sim N(0, 1 + \lambda^2 \sigma^2)}(Z'' < \lambda \mu + \beta) \\
= \Pr_{Y \sim N(0, 1)}(Y < \frac{\lambda \mu + \beta}{\sqrt{1 + \lambda^2 \sigma^2}}) \\
= \Phi\left(\frac{\lambda \mu + \beta}{\sqrt{1 + \lambda^2 \sigma^2}}\right)
\]

Side note: proving a special case of this result (with \(\beta = 0\)) is an exercise (Exercise 4.26) in Bishop and Nasrabadi (2006), who give a very complicated problem-specific solution calculating integrals (which can be found online). In contrast, this proof is strikingly simple and well-known in the Stack Exchange community (e.g., here and here). This is a reminder that often the “right” solution is simple, and even the best people can miss it.
Lemma G.35. An approximation of (32) is given by \( \hat{p}_{\text{test}} \in (0, 1)^M \) defined as

\[
[\hat{p}_{\text{test}}]_j = \text{sigmoid} \left( \frac{\pi \left( k([x_{\text{test}}]_j) - k([x_{\text{test}}]_j, x)A^{-1}k(x, [x_{\text{test}}]_j) \right) - 1/2 \times k([x_{\text{test}}]_j, x)(L(x) - \text{sigmoid}(y*))}{8} \right)
\]

where \( y^* \in \mathbb{R}^N \) is the mode of \( \text{posterior}(\cdot | L(x)) \) which necessarily satisfies \( y^* = (k(x) + \Sigma)(L(x) - \text{sigmoid}(y*)) \) and \( A = k(x) + \Sigma + \text{diag}(\text{sigmoid}(y^*) \odot (1 - \text{sigmoid}(y*)))^{-1} \).

Proof. Let \( a = 2L(x) - 1 \in \{ \pm 1 \}^N \) for symbolic convenience (as in SVMs). The posterior over logits is

\[
\text{posterior}(y | L(x)) \propto \left( \prod_{i=1}^{N} \text{sigmoid}(o_i y_i) \right) \times \mathcal{N}(0_N, k(x) + \Sigma)(y)
\]

We first apply the Laplace approximation (Appendix E) to obtain \( \text{posterior}(y | L(x)) \approx \mathcal{N}(y^*, -\nabla^2 l(y^*)^{-1}) \) where \( y^* \) is the mode and \( \nabla^2 l(y^*) \in \mathbb{R}^{N \times N} \) is the Hessian of the log posterior \( l(y) = \log \text{posterior}(y | L(x)) \) at \( y^* \).

Let \( g(y) = \log \text{likelihood}(L(x)|y) \) and \( h(y) = \log \text{prior}(y) \). The first- and second-order partial derivatives of \( g : \mathbb{R}^N \to \mathbb{R} \) are given by

\[
g(y) = \sum_{i=1}^{N} \log \text{sigmoid}(o_i y_i) \quad \quad \frac{\partial g(y)}{\partial y_i} = o_i (1 - \text{sigmoid}(o_i y_i)) = L(x_i) - \text{sigmoid}(y_i)
\]

\[
\frac{\partial^2 g(y)}{\partial y_i \partial y_j} = \begin{cases} -\text{sigmoid}(y_i)(1 - \text{sigmoid}(y_i)) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}
\]

Thus \( \nabla g(y) = L(x) - \text{sigmoid}(y) =: \delta \) and \( \nabla^2 g(y) = -W < 0 \) where \( W = \text{diag}(\text{sigmoid}(y) \odot (1 - \text{sigmoid}(y))). \)

Turning to \( h \) and denoting \( K = k(x) + \Sigma \),

\[
h(y) = -\frac{1}{2} y^T K^{-1} y + \text{constant} \quad \quad \nabla h(y) = -K^{-1} y \\
\nabla^2 h(y) = -K^{-1} < 0
\]

We have

\[
\nabla l(y) = \nabla g(y) + \nabla h(y) = \delta - K^{-1} y \\
\nabla^2 l(y) = \nabla^2 g(y) + \nabla^2 h(y) = -W - K^{-1} < 0
\]

which shows that the log posterior is strictly concave and the unique optimum is obtained at a point satisfying \( y^* = K\delta \). Denoting \( K' = k(x_{\text{test}}) \in \mathbb{R}^{M \times M} \) and \( R = k(x, x_{\text{test}}) \in \mathbb{R}^{N \times M} \), we have under the Laplace approximation (and using the stationary condition of \( y^* \))

\[
y \sim \mathcal{N}(K\delta, (W + K^{-1})^{-1}) \\
f_{\text{test}} | y \sim \mathcal{N}(R^T K^{-1} y, K' - R^T K^{-1} R)
\]

Then the marginal distribution is given by (9):

\[
f_{\text{test}} \sim \mathcal{N}(R^T \delta, K' - R^T (W^{-1} + K)^{-1} R)
\]

where the covariance matrix has been further simplified using the Woodbury identity \((K^{-1} + W)^{-1} = K - K(W^{-1} + K)^{-1} K \). Finally, we use an approximate closure of sigmoid under a Gaussian expectation in (28) which states that

\[
\mathbb{E}_{X \sim \mathcal{N}(\mu, \sigma^2)}[\text{sigmoid}(X)] \approx \text{sigmoid} \left( 1 + \frac{\pi \sigma^2}{8} \right)^{-1/2} \mu
\]

Applying this to individual test logits \( [f_{\text{test}}]_j \) in (97), we have \( (R_j \in \mathbb{R}^N \) denotes the j-th column of \( R) \)

\[
[p_{\text{test}}]_j = \mathbb{E}[\text{sigmoid}([f_{\text{test}}]_j)] \approx \text{sigmoid} \left( 1 + \frac{\pi \left( K'_{j,j} - R_j^T (W^{-1} + K)^{-1} R_j \right)}{8} \right)^{-1/2} \left( R_j^T \delta \right)
\]

□
Lemma G.36. Let \( z = (f_m, f) \in \mathbb{R}^{m \times N} \) be distributed as a Gaussian process (29) on inputs \((x_m, x) \in \mathcal{X}^{m+N}\) (which we omit from the notation since they are always observed) with some kernel \( k \). Let \( y \sim \mathcal{N}(f, \Sigma) \) denote the observed (noisy) labels of \( x \) where \( \Sigma \in \mathbb{R}^{N \times N}_+ \). The posterior distribution \( p(f_m|y) \) and the marginal log-likelihood (MLL) \( \log p(y) \) are given by

\[
p(f_m|y) = \mathcal{N}(\Lambda(x_m)^{-1}k(x_m, x)\Sigma^{-1}y, \Lambda(x_m)^{-1})(f_m)
\]

\[
\log p(y) = \log \mathcal{N}(0_N, \Sigma + Q(x_m))(y) - \frac{1}{2} \text{tr}(\Sigma^{-1}(k(x) - Q(x_m)))
\]

where \( \Lambda(x_m) = k(x_m)^{-1} + k(x_m)^{-1}k(x_m, x)\Sigma^{-1}k(x_m, x)k(x_m)^{-1} \) and \( Q(x_m) = k(x, x_m)k(x_m)^{-1}k(x_m, x) \).

Proof. Let us start from the definition of ELBO:

\[
\text{ELBO}(q) := \mathbb{E}_{z \sim q(y)} \left[ \log p(y|z) + \log \frac{p(z)}{q(z|y)} \right]
\]

The underlying “true” distributions are:

\[
\begin{align*}
\text{(prior)} & \quad p(z) = p(f_m) \times p(f|f_m) & \quad p(f_m) = \mathcal{N}(0_N, k(x_m))(f_m) \\
\text{(likelihood)} & \quad p(y|z) = p(y|f) = \mathcal{N}(f, \Sigma)(y) & \quad p(f|f_m) = \mathcal{N}(\alpha(x_m, f_m), k(x) - Q(x_m))(f)
\end{align*}
\]

\[
\text{(posterior)} & \quad p(z|y) = p(f_m|y) \times p(f|f_m)
\]

where \( \alpha(x_m, f_m) = k(x, x_m)k(x_m)^{-1}f_m \in \mathbb{R}^N \). The only unknown distribution is \( p(f_m|y) \). Our goal is to recover it (and compute the MLL) by optimizing the ELBO over an approximate posterior \( q(z|y) \). We can match its form to the true posterior \( q(z|y) := h(f_m, y) \times p(f|f_m) \) so that we only need to optimize over \( h \). Thus, viewing the ELBO as a function of \( h \) we have

\[
\text{ELBO}(h) = \mathbb{E}_{f_m \sim h(z|y)} \left[ \mathbb{E}_{f \sim p(\cdot|f_m)} \left[ \log p(y|f) \right] + \log \frac{p(f_m)}{h(f_m|y)} \right]
\]

Using (3),

\[
\mathbb{E}_{f \sim p(\cdot|f_m)} \left[ \log p(y|f) \right] = \log \mathcal{N}(\alpha(x_m, f_m), \Sigma)(y) - \frac{1}{2} \text{tr}(\Sigma^{-1}(k(x) - Q(x_m)))
\]

we eliminate the latent \( f \) in the ELBO, resulting in

\[
\text{ELBO}(h) = \mathbb{E}_{f_m \sim h(z|y)} \left[ \log \frac{\mathcal{N}(\alpha(x_m, f_m), \Sigma)(y) \times p(f_m)}{h(f_m|y)} \right] - \frac{1}{2} \text{tr}(\Sigma^{-1}(k(x) - Q(x_m)))
\]

(98)

The numerator is a joint distribution \( P(f_m, y) = p(f_m) \times p(y|f_m) \) with a Gaussian prior and likelihood, and the corresponding posterior can be computed by (10):

\[
P(f_m|y) = \mathcal{N}(\Lambda^{-1}k(x_m)^{-1}k(x_m, x)\Sigma^{-1}y, \Lambda^{-1})(f_m)
\]

where \( \Lambda = k(x_m)^{-1} + k(x_m)^{-1}k(x_m, x)\Sigma^{-1}k(x_m, x)k(x_m)^{-1} \). Since \( P(f_m, y) \propto P(f_m|y) \) when \( y \) is held fixed, for some \( C_y \) (constant in \( f_m \)) we have

\[
\text{ELBO}(h) = -\text{KL}(h(\cdot|y), P(\cdot|y)) + C_y
\]

This shows that the optimal \( h^*(f_m|y) \) is in fact \( P(f_m|y) \), which in turn gives \( p(f_m|y) = P(f_m|y) \) since \( h \) is not restricted. While we should be able to plug \( h^* \) in the ELBO to derive the MLL, we get the result faster by noting that the first term in (98) is actually a “mini-ELBO” associated with \( P(f_m, y) \) and \( h(f_m|y) \). Therefore, it attains its own maximum at \( \log P(y) \) where \( P(y) = \mathcal{N}(0_N, Q(x_m) + \Sigma)(y) \) is computed by (9). This shows

\[
\log p(y) = \log \mathcal{N}(0_N, Q(x_m) + \Sigma)(y) - \frac{1}{2} \text{tr}((\Sigma^{-1}(k(x) - Q(x_m)))
\]

---

8In the literature, an equivalent form is derived by convoluted algebraic manipulation. As a sanity check, we can convert to this form as follows. Let \( S = k(x_m)^{-1}k(x_m, x)\Sigma^{-1}k(x_m, x) \) and note that \( S = k(x_m)^{-1}k(x_m) \), thus \( \Lambda^{-1} = k(x_m)S^{-1}k(x_m) \) and \( \Lambda^{-1}k(x_m)^{-1}k(x_m, x)\Sigma^{-1}y = k(x_m)S^{-1}k(x_m, x)\Sigma^{-1}y \)

\[
\Lambda^{-1} = k(x_m)S^{-1}k(x_m)
\]

The right-hand sides become the mean and covariance proposed in Titsias (2009) with \( \Sigma = \sigma^2 I_{N \times N} \).

38
Individually Normal But Not Jointly Normal

This is an example from Wikipedia. Let $X \sim \mathcal{N}(0, 1)$ and, independently, $\epsilon \sim R$ where $R$ denotes the Rademacher distribution. Let $Y = \epsilon X$. By the symmetry of the distribution of $X$, we have $Y \sim \mathcal{N}(0, 1)$. More formally,

\[
\Pr(Y \leq x) = \Pr(\epsilon = 1) \Pr(X \leq x) + \Pr(\epsilon = -1) \Pr(X \geq -x) = \Pr(\epsilon = 1) \Pr(X \leq x) + \Pr(\epsilon = -1) \Pr(-X \leq x) = \frac{1}{2} \Pr(X \leq x) + \frac{1}{2} \Pr(X \leq x) = \Pr(X \leq x)
\]

Let $Z = X + Y$. Then $Z = 0$ with probability $\frac{1}{2}$ and $Z = 2X$ with probability $\frac{1}{2}$, so

\[
\Pr(Z = z) = \frac{1}{2} \left( [z = 0] + \mathcal{N}(0, 1) \left( \frac{z}{2} \right) \right)
\]

which is not a normal distribution. Then by definition 4, $(X, Y) \in \mathbb{R}^2$ is not normally distributed. Thus $X$ and $Y$ are not jointly normal, even though they are individually normal.

**Mutual information.** $X$ and $Y$ are uncorrelated. More formally,

\[
\text{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X] \mathbf{E}[Y] = \mathbf{E}[\epsilon X^2] = \mathbf{E}[\epsilon] \mathbf{E}[X^2] = 0
\]

Thus $\text{cor}(X, Y) = 0$. But $X$ and $Y$ are not independent. Specifically, $\Pr(Y = x | X = x) = \frac{1}{2}$ is not equal to $\Pr(Y = x) = \mathcal{N}(0, 1)(x)$ for any $x \in \mathbb{R}$. This illustrates the limitation of linear correlation. On the other hand, the mutual information between $X$ and $Y$ is positive:

\[
I(X, Y) = H(X) - H(X|Y) = H(X) - \log(2) = \log \sqrt{\pi e \frac{1}{2}} \approx 0.73
\]