

Problem 1. Markov Chain Monte Carlo

1.

① Show that x follows F .

Since we have $F^{-1}(U) \leq x$

$$X \sim F^{-1}(U) \Leftrightarrow F \circ F^{-1}(U) \leq F(x)$$

$$\Leftrightarrow U \leq F(x)$$

Thus, χ follows F.

② The drawback of this method?

F^{-1} should be explicitly valid to use this method.

2.

- ① Given that k_1 and k_2 have $p(\cdot)$ as stationary density,
we have

$$\underbrace{\int p(x) \cdot k_1(x, y) dx = p(y)}_{\text{(*)}} \quad \text{and}$$

$$\underbrace{\int p(x) \cdot k_2(x, y) dx = p(y).}_{\text{(**)}}$$

- ② Show that $k_1 \circ k_2$ has $p(\cdot)$ as stationary density.

$$\begin{aligned} & \int p(x) \cdot \underbrace{k_1 \circ k_2(x, z)}_{\text{by (1)}} dx \\ &= \int p(x) \underbrace{\int k_2(x, y) k_1(y, z) dy dx}_{\text{by (**)}} \\ &= \int p(y) k_1(y, z) dy. \\ &= p(z). \end{aligned}$$

$\therefore k_1 \circ k_2$ has $p(\cdot)$ as stationary density.

③. Show that $\lambda k_1 + (1-\lambda) k_2$ for any $\lambda \in [0,1]$ has $p(\cdot)$ as stationary density.

$$\int p(x) (\lambda k_1(x,y) + (1-\lambda) k_2(x,y)) dx$$

$$= \lambda \underbrace{\int p(x) k_1(x,y) dx}_{\text{by } \oplus} + (1-\lambda) \underbrace{\int p(x) k_2(x,y) dx}_{\text{by } \oplus}$$

$$= \lambda \underbrace{p(y)}_{\text{by } \oplus} + (1-\lambda) \cdot \underbrace{p(y)}_{\text{by } \oplus}$$

$$= p(y)$$

$\therefore \lambda k_1 + (1-\lambda) k_2$ has $p(\cdot)$ as stationary density.

3.

We can construct the transition kernel K as :

$$K(s,t) = g(t|s) A(s,t) + (1 - r(s)) \delta_s(t),$$

where $A(s,t) = \min\left(1, \frac{\tilde{p}(t) g(s|t)}{\tilde{p}(s) g(t|s)}\right) = \min\left(1, \frac{p(t) g(s|t)}{p(s) g(t|s)}\right)$ is the acceptance probability of a move from state s to state t ,

$$r(s) = \int g(t|s) A(s,t) dt, \text{ and } \delta_s(t) \text{ is the Dirac mass at } s.$$

We want to show that p satisfies detailed balance for all s and t .

That is, we want to show that

$$p(s) K(s,t) \stackrel{?}{=} p(t) K(t,s).$$

$$\textcircled{1} \text{ LHS} = p(s) K(s,t)$$

$$= p(s) [g(t|s) A(s,t) + (1 - r(s)) \delta_s(t)]$$

$$= p(s) g(t|s) A(s,t) + p(s) (1 - r(s)) \delta_s(t)$$

$$= (p(s) g(t|s) \text{ or } p(t) g(s|t)) + p(s) (1 - r(s)) \delta_s(t)$$

$$\begin{array}{ccc} \uparrow & \uparrow \\ \text{When } A(s,t)=1 & \text{When } A(s,t) \neq 1 \end{array}$$

$$\textcircled{2} \quad \text{RHS} = p(t) k(t,s)$$

$$= p(t) [g(s|t) A(t,s) + (1-r(t)) \delta_t(s)]$$

$$= \left(\begin{array}{l} p(t) g(s|t) \text{ or } p(s) g(t|s) \\ \text{when } A(t,s)=1 \qquad \qquad \text{when } A(t,s) \neq 1 \end{array} \right) + p(t) (1-r(t)) \delta_t(s)$$

$$\textcircled{3} \quad \text{LHS} - \text{RHS}$$

$$= \left(\begin{array}{l} p(s) g(t|s) \text{ or } p(t) g(s|t) \\ \text{when } A(s,t)=1 \qquad \text{when } A(s,t) \neq 1 \end{array} \right) - \left(\begin{array}{l} p(t) g(s|t) \text{ or } p(s) g(t|s) \\ \text{when } A(t,s)=1 \qquad \text{when } A(t,s) \neq 1 \end{array} \right)$$

$$+ p(s) (1-r(s)) \delta_s(t) - p(t) (1-r(t)) \delta_t(s)$$

if $s=t$, this term is 0.

$$\text{otherwise, } p(s) (1-r(s)) \delta_s(t) = p(t) (1-r(t)) \delta_t(s) = 0$$

thus this term is 0 as well.

$$= 0 + 0 = 0.$$

$\therefore \text{LHS} = \text{RHS}$, p satisfies detailed balance.

Thus, p is the stationary distribution.

4.

The transition kernel of Gibbs chain is

$$K(x, x') = p_1(x'_1 | x_2, \dots, x_d) \times p_2(x'_2 | x'_1, x_3, \dots, x_d) \\ \times \dots \times p_d(x'_d | x'_1, \dots, x'_{d-1}).$$

To show that $p(x)$ is the stationary distribution of the Markov chain, we want to show that

$$\int K(x, x') p(x) dx = p(x').$$

Let $\hat{p}_i(\cdot)$ be the marginal over all variables except for x_i .

$$\int K(x, x') p(x) dx$$

$$= \underbrace{\int p_1(x'_1 | x_2, \dots, x_d)}_{\hat{p}_1(x_2, \dots, x_d)} \cdot p_2(x'_2 | x'_1, x_3, \dots, x_d) \\ \times \dots \times p_d(x'_d | x'_1, \dots, x'_{d-1}) \\ \times \underbrace{\hat{p}_1(x_2, \dots, x_d)}_{\hat{p}(x_1 | x_2, \dots, x_d)} \underbrace{\boxed{p(x_1 | x_2, \dots, x_d) dx_1}}_{\dots} \dots dx_d$$

$$= \int p_2(x'_2 | x'_1, x_3, \dots, x_d) \times \dots \times p_d(x'_d | x'_1, \dots, x'_{d-1}) \\ \times \underbrace{p(x'_1, x_2, \dots, x_d)}_{\hat{p}(x_1, x_2, \dots, x_d)} dx_2 \dots dx_d$$

$$\begin{aligned}
 &= \int P_2(x_2' | x_1', x_3, \dots, x_d) \times \cdots \times P_d(x_d' | x_1', \dots, x_{d-1}) \\
 &\quad \times \underbrace{\hat{P}_2(x_1', x_3, \dots, x_d)}_{P(x_2 | x_1', x_3, \dots, x_d)} \underbrace{P(x_d | x_1', x_3, \dots, x_d)}_{d x_2 \cdots d x_d} \\
 &= \cdots \\
 &= p(x_1', \dots, x_d') \\
 &= p(\mathbf{x}').
 \end{aligned}$$

Thus, $p(\mathbf{x})$ is the stationary distribution.

Problem 2 Box-Muller method

I refer David Barber's text book (Bayesian and Machine Learning).
(Change of Variable)

For a univariate continuous random variable x with distribution $p(x)$, the transformation $y=f(x)$, where $f(x)$ is a monotonic function, has distribution

$$p(y) = p(x) \left(\frac{df}{dx} \right)^{-1}, \quad x = f^{-1}(y).$$

Let's denote $\mathbf{y} = (y_1, y_2)$ and $\mathbf{x} = (x_1, x_2)$.

For multivariate \mathbf{x} and bijection $f(\mathbf{x})$, then $\mathbf{y} = f(\mathbf{x})$ has distribution

$$p(\mathbf{y}) = p(\mathbf{x} = f^{-1}(\mathbf{y})) \left| \det \left(\frac{\partial f}{\partial \mathbf{x}} \right) \right|^{-1}$$

where the Jacobian matrix has elements

$$\left[\frac{\partial f}{\partial \mathbf{x}} \right]_{ij} = \frac{\partial f_i(\mathbf{x})}{\partial x_j}.$$

$$\textcircled{1} \quad \left| \det \frac{\partial f}{\partial x} \right| = \left| \det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} \right|$$

$$\cdot \frac{\partial y_1}{\partial x_1} = \frac{\partial}{\partial x_1} (\sqrt{-2 \log x_1} \cos 2\pi x_2)$$

$$= \frac{-2 \cdot \frac{1}{x_1}}{2\sqrt{-2 \log x_1}} \cdot \cos 2\pi x_2 = \frac{-1}{\sqrt{-2 \log x_1}} \cdot \frac{\cos 2\pi x_2}{x_1}$$

$$\cdot \frac{\partial y_1}{\partial x_2} = \frac{\partial}{\partial x_2} (\sqrt{-2 \log x_1} \cos 2\pi x_2)$$

$$= \sqrt{-2 \log x_1} \cdot (-\sin(2\pi x_2)) \cdot 2\pi = -2\pi \cdot \sqrt{-2 \log x_1} \cdot \sin 2\pi x_2$$

$$\cdot \frac{\partial y_2}{\partial x_1} = \frac{\partial}{\partial x_1} (\sqrt{-2 \log x_1} \sin 2\pi x_2)$$

$$= \frac{-2 \cdot \frac{1}{x_1}}{2\sqrt{-2 \log x_1}} \sin 2\pi x_2 = \frac{-1}{\sqrt{-2 \log x_1}} \cdot \frac{\sin 2\pi x_2}{x_1}$$

$$\cdot \frac{\partial y_2}{\partial x_2} = \frac{\partial}{\partial x_2} (\sqrt{-2 \log x_1} \sin 2\pi x_2)$$

$$= \sqrt{-2 \log x_1} \cdot \cos 2\pi x_2 \cdot 2\pi$$

$$\therefore \left| \det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} \right| = \left| -\frac{2\pi}{x_1} \cos^2 2\pi x_2 - \frac{2\pi}{x_1} \sin^2 2\pi x_2 \right| = \frac{2\pi}{x_1}$$

$$\textcircled{2} \quad y_1 = \sqrt{-2 \log x_1} \cos 2\pi x_2, \quad y_2 = \sqrt{-2 \log x_1} \sin 2\pi x_2$$

↓

$$\cos 2\pi x_2 = \frac{y_1}{\sqrt{-2 \log x_1}} \quad \sin 2\pi x_2 = \frac{y_2}{\sqrt{-2 \log x_1}}$$

$$\cos^2 2\pi x_2 + \sin^2 2\pi x_2 = 1 = \frac{y_1^2 + y_2^2}{-2 \log x_1}$$

$$\therefore x_1 = \exp \left(-\frac{y_1^2 + y_2^2}{2} \right)$$

$$\textcircled{3} \quad p(y) = p(\mathbf{x} = f^{-1}(y)) \left| \det \left(\frac{\partial f}{\partial \mathbf{x}} \right) \right|^{-1} \leftarrow \begin{array}{l} \text{by change of} \\ \text{variable} \end{array}$$

$$= p(\mathbf{x} = f^{-1}(y)) \cdot \frac{x_1}{2\pi} \leftarrow \text{by } \textcircled{1}$$

$$= p(\mathbf{x} = f^{-1}(y)) \cdot \frac{1}{2\pi} \cdot e^{-\frac{1}{2}(y_1^2 + y_2^2)} \leftarrow \text{by } \textcircled{2}$$

$$= 1 \cdot \frac{1}{2\pi} \cdot e^{-\frac{1}{2}(y_1^2 + y_2^2)} \leftarrow p(x_1) = p(x_2) = 1$$

$$= \mathcal{N}(y_1 | 0, 1) \mathcal{N}(y_2 | 0, 1)$$

$$\therefore \text{We arrive at } p(y_1, y_2) = \mathcal{N}(y_1 | 0, 1) \mathcal{N}(y_2 | 0, 1)$$

Problem 3

We know that all neighbors of any white variables are all black, and also that all neighbors of any black variables are all white. When we condition on all black variables, the white variables become independent. Similarly, when we condition on all black variables, the black variables become independent.

Thus, $P(b_1, b_2, \dots | w_1, w_2, \dots) = P(b_1 | w_1, \dots) \cdot P(b_2 | w_1, \dots) \cdots$
 and $P(w_1, w_2, \dots | b_1, b_2, \dots) = P(w_1 | b_1, \dots) \cdot P(w_2 | b_1, \dots) \cdots$.

In Gibbs Sampling procedure, we can draw white samples from their black neighbors conditioned on the white variables by using the above joint distribution. Also, we can draw black samples from their white neighbors conditioned on the black variables by using the above joint distribution.

Problem 4

The code is in Problem-4/symptomDisease.m. In the code, I used 500 burn-in, and 10000 sub-sampling. The vector computed is:

=====

Columns 1 through 10

0.0029 0.9979 0.0214 1.0000 0.6472 0.0115 0.0149 0.0008 0.0053 0.9999

Columns 11 through 20

0.0051 1.0000 1.0000 1.0000 0.9907 0.9677 0.9898 0.9055 0.9998 0.9933

Columns 21 through 30

0.0938 0.7473 1.0000 0.9935 0.0056 0.9996 0.0173 0 0.9986 0

Columns 31 through 40

0 0.0996 0.0115 1.0000 0 0.0031 0.9938 0.0015 0 0.0002

Columns 41 through 50

0.9927 0.9985 1.0000 0.9997 0.0001 0.0167 0.9959 0.9966 0.0002 0.9993

Problem 5

1) Derive $p(d \mid s, D) = \int_{lw, lb, lp} p(d \mid s, D, lw, lb, lp) p(lw, lb, lp \mid D)$

$$\begin{aligned} p(d \mid s, D) &= \int_{lw, lb, lp} p(d \mid s, D, lw, lb, lp) p(lw, lb, lp \mid D) \\ &= \int_{lw, lb, lp} p(d \mid s, lw, lb, lp) p(lw, lb, lp \mid D) \end{aligned}$$

↑

Since the prediction is independent to D .

2). Explain how to estimate $p(d_i=1 \mid s, D)$ using Sampling.

We can use Gibbs sampling. By fixing lb and lp , we can perform Gibbs sampling on $(d \mid s, D)$. By fixing lw, lp , we can perform Gibbs sampling on $(d \mid s, D)$. By fixing lw, lb , we can perform Gibbs sampling on $(d \mid s, D)$.

Through this controlled Gibbs sampling, we can estimate $p(d_i=1 \mid s, D)$.

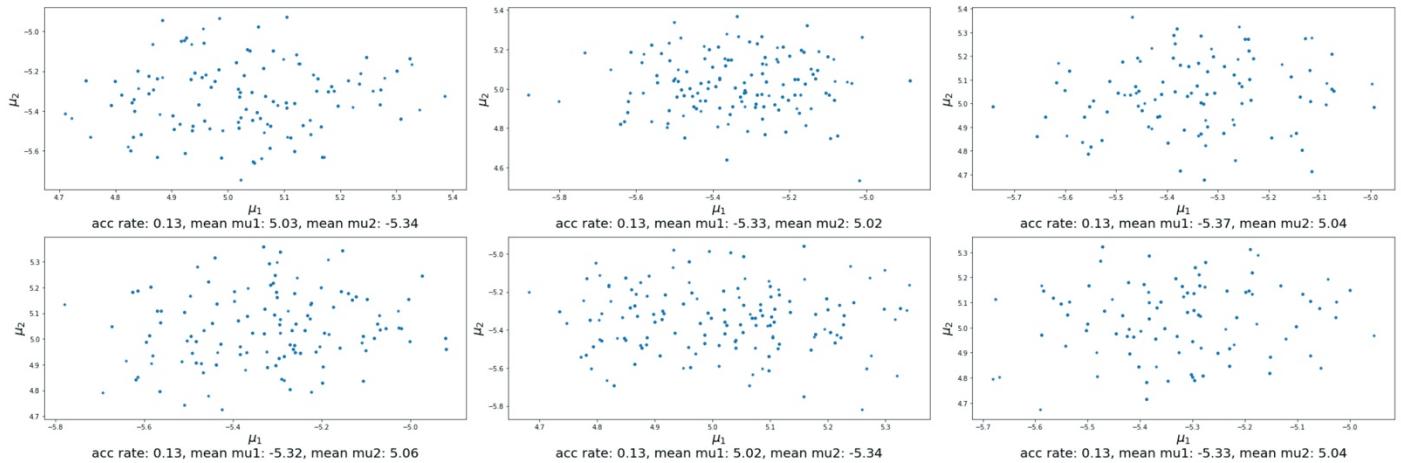
Problem 6

The code is in Problem-6/problem_6.ipynb. The pdf version of the notebook is Problem-6/problem_6.pdf.

a) See problem_6.ipynb, Problem a).

b)

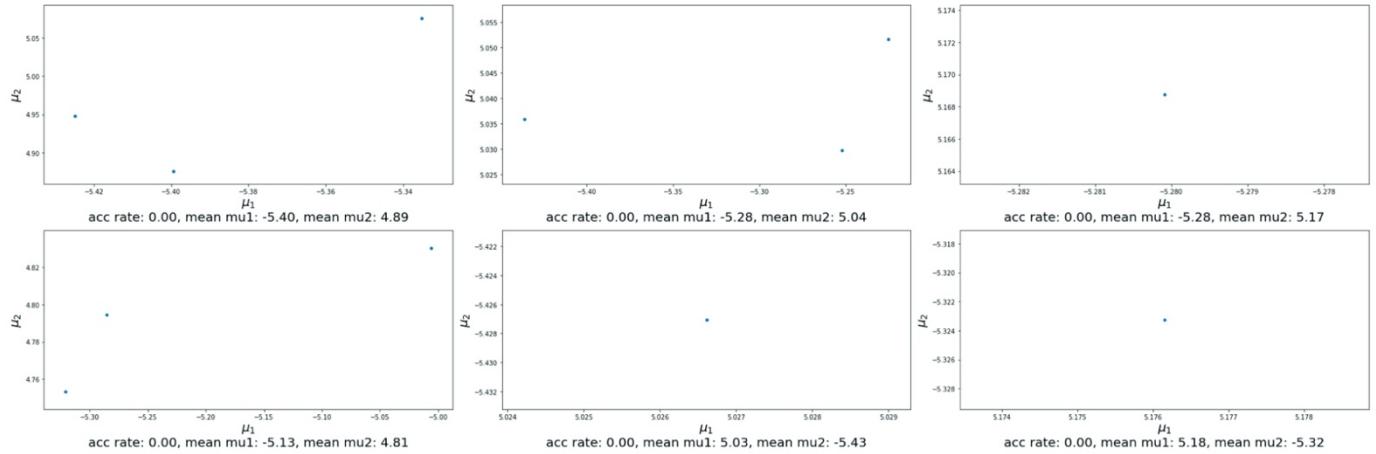
When sigma = 0.5, the plot of the last 1000 mu1 and mu2 is:



The estimated mean of mu1 and mu2, and the acceptance rate are:

mean_mu1	mean_mu2	acceptance rate
5.025064	-5.340499	0.131000
-5.332442	5.021185	0.131091
-5.365892	5.035718	0.130727
-5.316841	5.056698	0.127909
5.019107	-5.335424	0.134545
-5.333710	5.039708	0.127909

When sigma = 5, the plot of the last 1000 mu1 and mu2 is:



The estimated mean of mu1 and mu2, and the acceptance rate are:

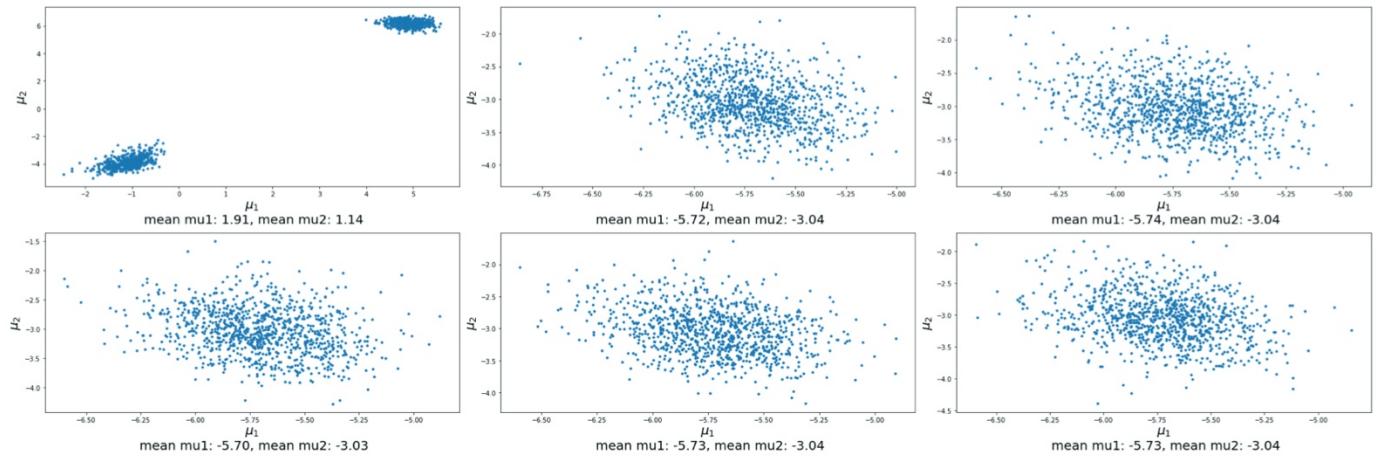
mean_mu1	mean_mu2	acceptance rate
-5.401915	4.894947	0.002273
-5.277091	5.037582	0.001455
-5.280091	5.168751	0.002182
-5.128976	4.807459	0.003273
5.026610	-5.427057	0.001909
5.176153	-5.323255	0.002182

The differences between sigma = 0.5 and sigma = 5 are:

- When sigma = 5, the acceptance rate is lower, compared to when sigma = 0.5.
- When sigma = 0.5, the mu1 and mu2 are more scattered, compared to when sigma = 5.

c)

The plot of the last 1000 mu1 and mu2 is:



The estimated mean of mu1 and mu2 are:

mean_mu1	mean_mu2
1.912066	1.142613
-5.724099	-3.042464
-5.738514	-3.035543
-5.704386	-3.032615
-5.725242	-3.040966
-5.728205	-3.044541

Problem 7

Run Problem-7/KL.m. We can get the value of the minimal KL Divergence for the optimal q as 0.0159647.

Problem 8

Run Problem-8/BPMF.m.

1. The result of loopy belief propagation

i	p_i	q_i
1	0.0364	0.9636
2	0.7064	0.2936
3	0.4510	0.5490
4	0.8302	0.1698

The optimal fully factorized approximation is

$$\prod_{i=1}^4 q_i^{BP} = 0.9636 * 0.2936 * 0.5490 * 0.1698 = 0.0264$$

2. The result of variational mean-field equation

i	p_i	q_i
1	0.0021	0.9979
2	0.9216	0.0784
3	0.5668	0.4332
4	0.8779	0.1221

The optimal fully factorized approximation is

$$\prod_{i=1}^4 q_i^{MF} = 0.9979 * 0.0784 * 0.4332 * 0.1221 = 0.0041$$

3. The marginals of exact inference

i	p_i	q_i
1	0.0325	0.9675
2	0.7104	0.2896
3	0.4509	0.5491
4	0.8292	0.1708

4. Mean expected deviation in the marginals:

$$BP = 0.00218917$$

$$MF = 0.101594$$

Belief propagation works better since it has lower mean error.

Problem 9.

1.

Let's denote z be the normalization coefficient for $r(x)$. Then,

$$r(x) = \frac{p(x)f(x)}{z}$$

By construction, $J = \log z$.

Taking $\text{KL}(r|g)$ gives

$$\begin{aligned} \log z &= J \\ &\geq -\langle \log g(x) \rangle_{g(x)} + \langle \log p(x) \rangle_{g(x)} + \langle \log f(x) \rangle_{g(x)} \\ &= -\text{KL}(g(x) \mid p(x)) + \langle \log f(x) \rangle_{g(x)} \end{aligned}$$

2.

$$J \geq -\text{KL}(g(x) \mid p(x)) + \langle \log f(x) \rangle_{g(x)} \quad \leftarrow \text{by 1}$$

$$\begin{aligned} &= -\text{KL}(g(x) \mid p(x)) + \underbrace{\langle \log f(x) \rangle_{g(x)} - \langle \log g(x) \rangle_{g(x)}}_{\downarrow} + \langle \log g(x) \rangle_{g(x)} \\ &= -\text{KL}(g(x) \mid p(x)) - \underbrace{\text{KL}(g(x) \mid f(x))}_{\downarrow} + \langle \log g(x) \rangle_{g(x)} \\ &= -\text{KL}(g(x) \mid p(x)) - \text{KL}(g(x) \mid f(x)) - H(g(x)). \end{aligned}$$