Métodos Numéricos Probabilísticos

Roberto Díaz Morales

DTSC, Universidad Carlos III de Madrid

Índice

Introducción

- Una de las tareas del aprendizaje máquina consiste en la inferencia sobre datos complejos.
- Dicha tarea incluye el desarrollo y análisis de métods numéricos basados en teoría probabilística.
- Muchas de estas tareas se pueden ver como problemas de aprendizaje.
- Dos de las tareas más importantes son:
  - Cuadratura
  - Muestreo

Cuadratura

- Muchos modelos complejos requieren integrales computacionalmente intratables, con lo que han de ser aproximadas.
- En particular, muchos problemas de inferencia requieren integrar sobre funciones de probabilidad.

$$Z = \langle \ell \rangle = \int \ell(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

  - p(x) podría ser un “posterior” y f(x) las etiquetas de nuestras muestras.
  - p(x) podría ser un “prior” y f(x) una verosimilitud.
  - …

- Es complicado cuando se trabaja con grandes conjuntos de datos donde evaluar la verosimilitud sobre todo el conjunto de datos es muy costoso computacionalmente.

Índice

Estimador de Monte Carlo

- Si p(x) es una función de densidad de probabilidad y podemos obtener muestras de ella tenemos la aproximación de Monte Carlo:

$$\Phi = \langle \phi(\mathbf{x}) \rangle \equiv \int P(\mathbf{x}) \phi(\mathbf{x}) \, \mathrm{d}^N \mathbf{x}$$

$$\hat{\Phi} \equiv \frac{1}{R} \sum_r \phi(\mathbf{x}^{(r)}) \qquad \sigma^2 = \int \mathrm{d}^N \mathbf{x} \; P(\mathbf{x}) (\phi(\mathbf{x}) - \Phi)^2$$

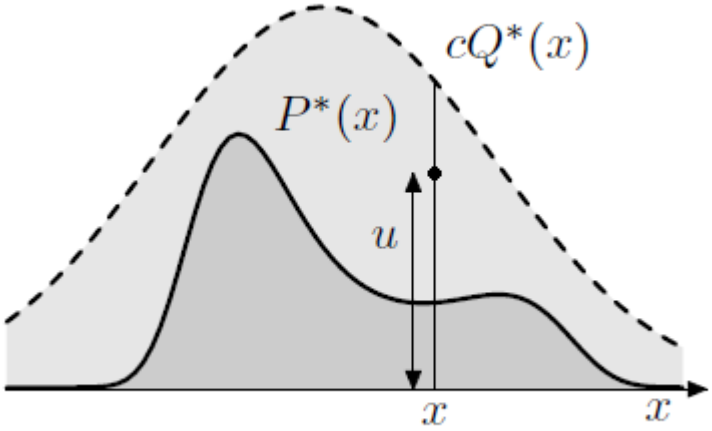
- Su varianza decae en un orden O(1/R)
- La principal objección es que tener muestras que representen bien P(x) no garantiza que representen bien Φ(x)

Índice

1. Introducción
2. Cuadratura
3. Muestreo
4. Cuadratura Bayesiana
5. Aprendizaje activo de la evidencia usando cuadratura bayesiana.
6. Conclusiones

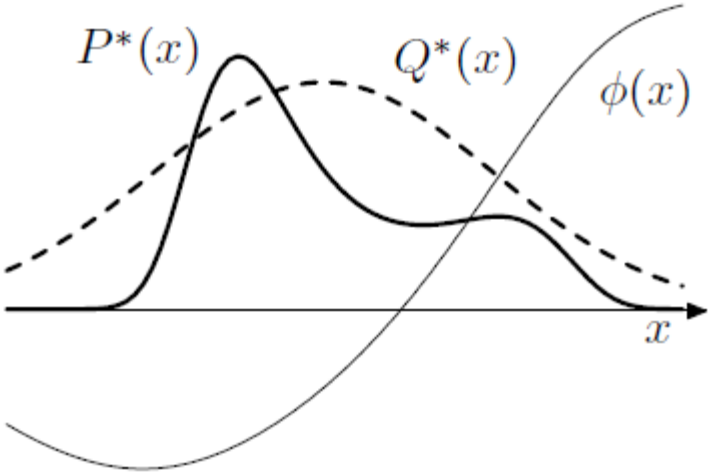
Métodos de muestreo

- Existen multitud de métodos de muestreo cuando podemos evaluar p(x):
  - Rejection Sampling:
    - Se conoce cQ(x)>P(x)
    - Se genera x_i muestra de cQ(x)
    - Se genera u~U(0,cQ(x_i))
    - Si u<P(x) guardamos x_i
    - Si u>P(x) descartamos x_i
    - Eficiencia según el parecido de cQ(x) con P(x)



Métodos de muestreo

- Importance Sampling:



$$w_r \equiv \frac{P^*(x^{(r)})}{Q^*(x^{(r)})} \qquad \hat{\Phi} \equiv \frac{\sum_r w_r \phi(x^{(r)})}{\sum_r w_r}$$

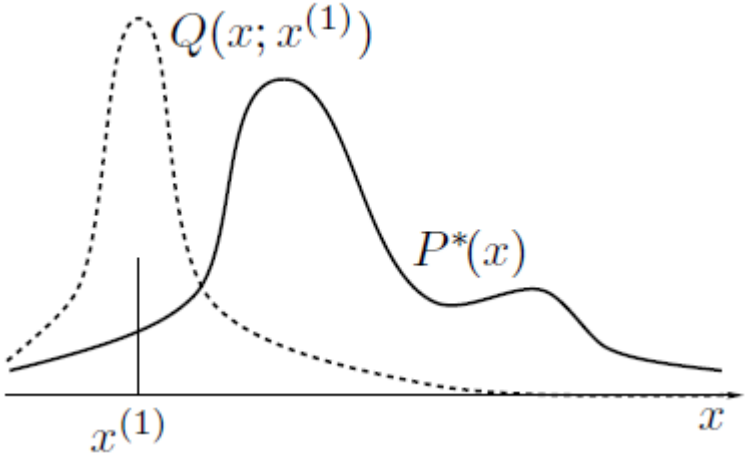
Métodos de muestreo MCMC

- Markov Chain Monte Carlo:
  - Son técnicas de muestreo en los que la siguiente muestra depende de la muestra actual (estado) y de unas probabilidades de transición.

- Metropolis-Hasting:
  - Se utiliza una función Q(x) que depende de la muestra actual.
  - Se evalua a:

$$a = \frac{P^*(x')}{P^*(x^{(t)})} \frac{Q(x^{(t)}; x')}{Q(x'; x^{(t)})}$$

  - Si a>1 la muestra se acepta.
  - Si a<1 se acepta con probabilidad a.
  - Si se acepta es el nuevo estado.
  - Se debe dejar correr $T \simeq (L/\epsilon)^2$



Métodos de muestreo MCMC

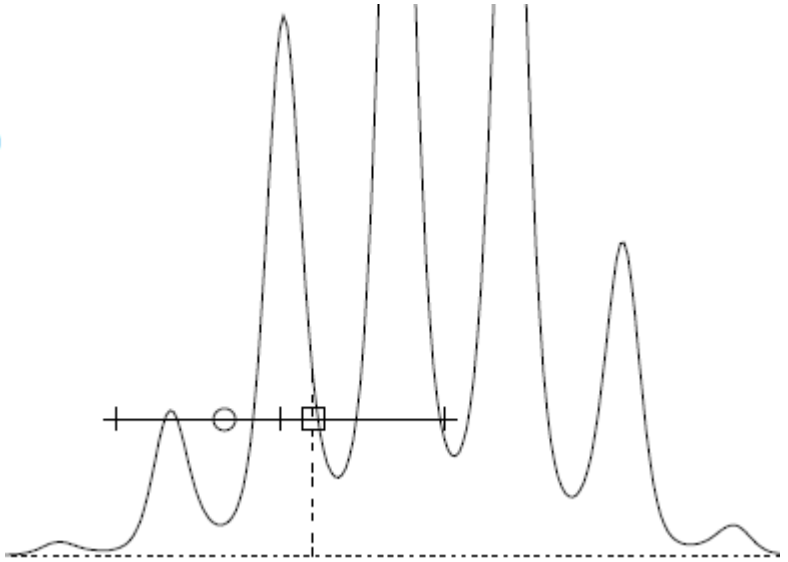
- Gibbs-Sampling:
  - La siguiente muestra se obtiene con la distribución conjunta para cada una de las dimensiones.

$$
\begin{aligned}
x_1^{(t+1)} &\sim P(x_1 \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\
x_2^{(t+1)} &\sim P(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\
x_3^{(t+1)} &\sim P(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_K^{(t)})
\end{aligned}
$$

Métodos de muestreo MCMC

- Slice-Sampling:

1: evaluate $P^*(x)$
2: draw a vertical coordinate $u' \sim \text{Uniform}(0, P^*(x))$
3: create a horizontal interval (x_l, x_r) enclosing x
4: loop $\{$
5: draw $x' \sim \text{Uniform}(x_l, x_r)$
6: evaluate $P^*(x')$
7: if $P^*(x') > u'$ break out of loop 4-9
8: else modify the interval (x_l, x_r)
9: $\}$



Índice

Cuadratura Bayesiana

- Rasmussen, C. E., & Ghahramani, Z.
  Bayesian monte carlo. *Advances in neural information processing systems*, *15*, 489-496.

- Dado un conjunto de muestras D y realizando inferencia sobre f con D, la media sobre funciones es la esperanza de f(x) media.

$$\mathcal{D} = \{(x^{(i)}, f(x^{(i)}))|i = 1 \ldots n\}$$

$$E_{f|\mathcal{D}}[\bar{f}_p] = \iint f(x)p(x)dx \; p(f|\mathcal{D})df$$

$$V_{f|\mathcal{D}}[\bar{f}_p] = \int \left[\int f(x)p(x)dx - \int \bar{f}(x')p(x')dx' \right]^2 p(f|\mathcal{D})df$$

Cuadratura Bayesiana

- Se agrupan los términos para un GP:

$$E_{f|\mathcal{D}}[\bar{f}_p] = \iint f(x)p(x)dx\; p(f|\mathcal{D})df$$
$$= \int \Big[\int f(x)p(f|\mathcal{D})df \Big] p(x)dx = \int \bar{f}_{\mathcal{D}}(x)p(x)dx$$

$$V_{f|\mathcal{D}}[\bar{f}_p] = \int \Big[\int f(x)p(x)dx - \int \bar{f}(x')p(x')dx' \Big]^2 p(f|\mathcal{D})df$$
$$= \iiint \big[f(x) - \bar{f}(x)\big] \big[f(x') - \bar{f}(x')\big] p(f|\mathcal{D})df p(x)p(x')dxdx'$$
$$= \iint \mathrm{Cov}_{\mathcal{D}}\big(f(x), f(x')\big)p(x)p(x')dxdx',$$

$$\bar{f}_{\mathcal{D}}(x) = k(x,\mathbf{x})K^{-1}\mathbf{f}, \text{ and } \mathrm{Cov}_{\mathcal{D}}\big(f(x), f(x')\big) = k(x,x') - k(x,\mathbf{x})K^{-1}k(\mathbf{x},x')$$

Cuadratura Bayesiana

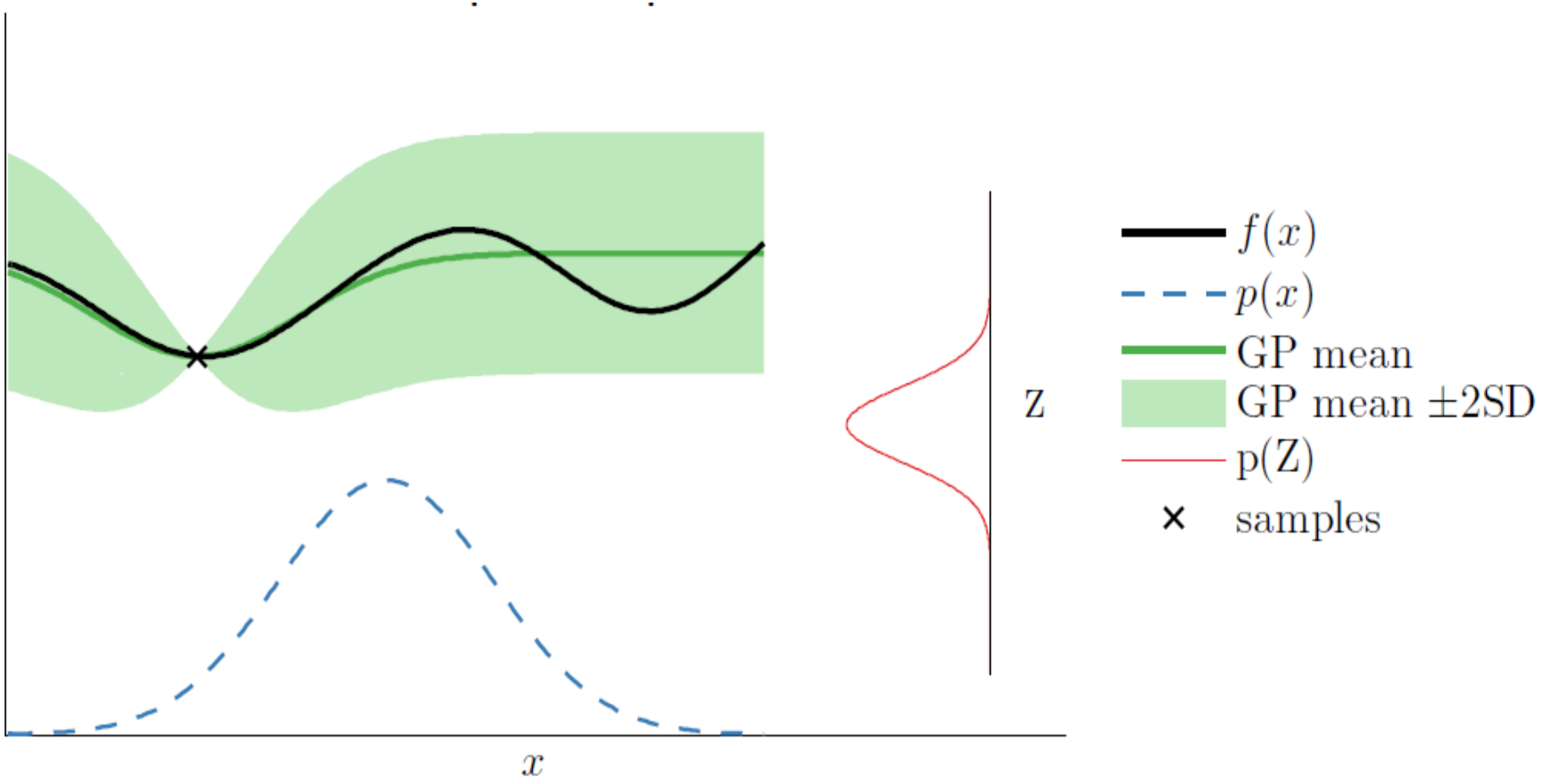
- En el caso general, introducir la formulación de los GP en la integral lleva a expresiones que son dificiles de evaluar, pero hay casos especiales:
  - Si p(x) y la función de covarianza son ambas gaussianas, se obtienen expresiones analíticas (utilizando la cuadratura Bayes-Hermite):

$$E_{f|\mathcal{D}}[\bar{f}_p] = z^\top K^{-1}\mathbf{f}$$
$$V_{f|\mathcal{D}}[\bar{f}_p] = w_0 \left|2A^{-1}B + I\right|^{-1/2} - z^\top K^{-1} z$$

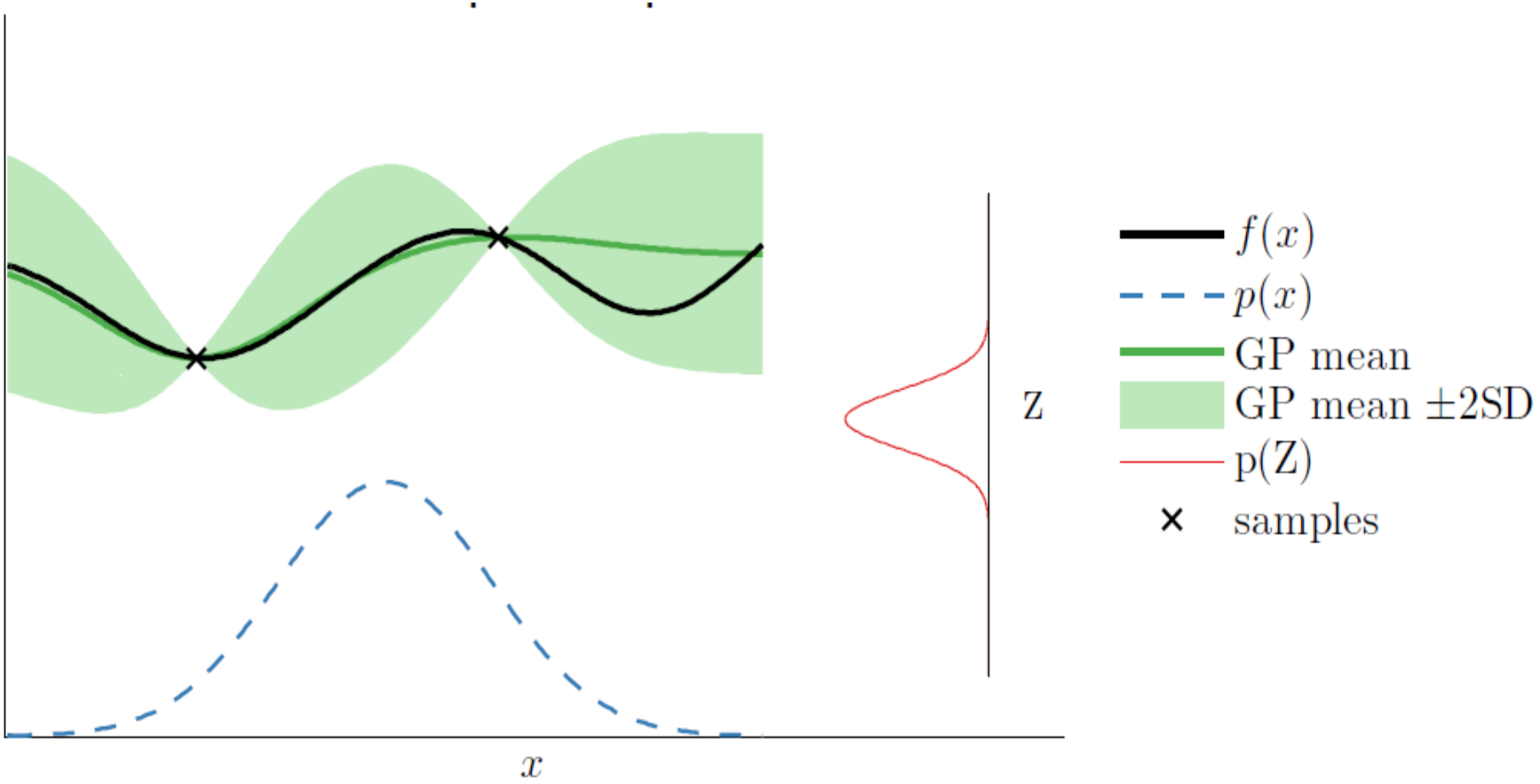
$$z = w_0|A^{-1}B+I|^{-1/2}\exp[-0.5(a-b)^\top(A+B)^{-1}(a-b)]$$
$$p(x) = \mathcal{N}(b, B)$$
$$a_i = x^{(i)}$$

$$K_{pq} = \mathrm{Cov}(f(x^{(p)}), f(x^{(q)})) = w_0 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}(x_d^{(p)} - x_d^{(q)})^2/w_d^2\right)$$

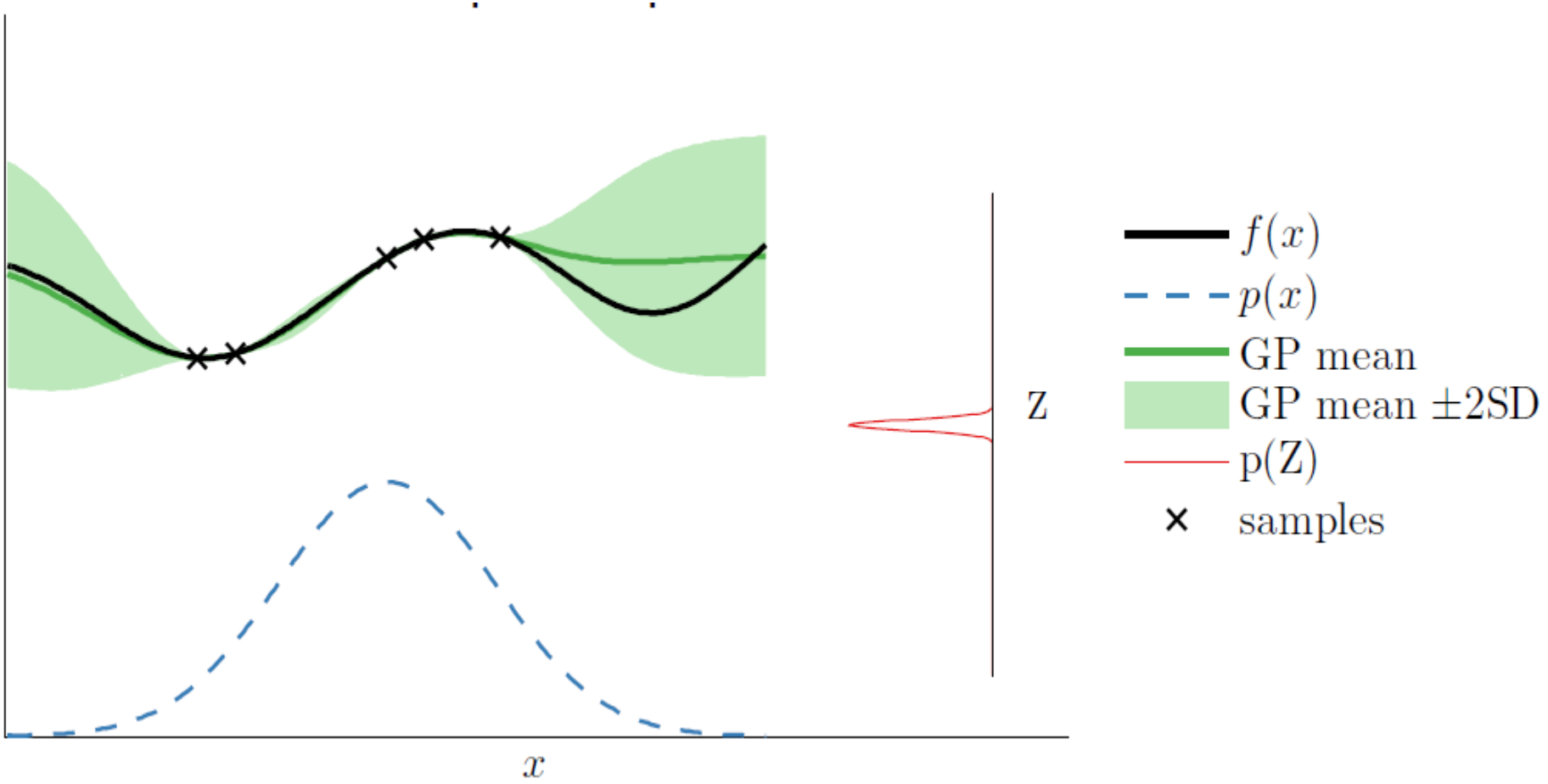
Cuadratura Bayesiana



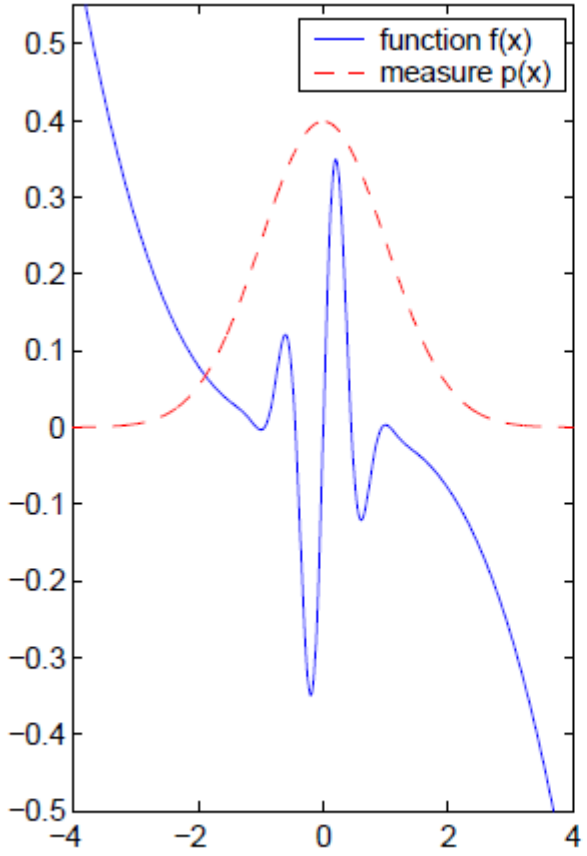
Cuadratura Bayesiana

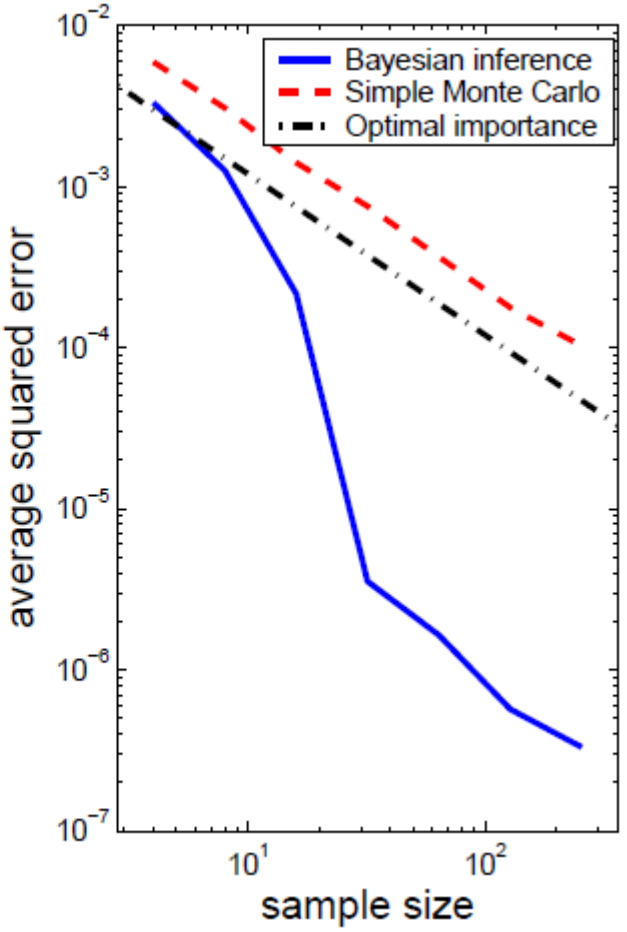


Cuadratura Bayesiana



Cuadratura Bayesiana





Índice

1. Introducción
2. Cuadratura
3. Muestreo
4. Cuadratura Bayesiana
5. Aprendizaje activo de la evidencia usando cuadratura bayesiana.
6. Conclusiones

Aprendizaje Activo de Evidencia

- Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S. J., & Ghahramani, Z. Active Learning of Model Evidence Using Bayesian Quadrature. NIPS 2012

- Se desea realizar una integral sobre una verosimilitud no negativa:

$$Z = \langle \ell \rangle = \int \ell(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$m(Z|\log \ell_s) = \int \Big(\int \exp\big(\log \ell(x)\big) p(x)\ \mathrm{d}x \Big) \mathcal{N}\big(\log \ell; m_{\log \ell|s}, C_{\log \ell|s}\big)\ \mathrm{d}\log \ell$$

Aprendizaje Activo de Evidencia

- Para poder tratar la integral, "lineariza" el problema.

$$\exp\big(\log \ell(x)\big) \simeq \exp\big(\log \ell_0(x)\big) + \exp\big(\log \ell_0(x)\big)\big(\log \ell(x) - \log \ell_0(x)\big)$$

- Quedando como problema a resolver:

$$m(Z|\log \ell_s) \simeq m(Z|\log \ell_0, \log \ell_s) := \int \ell_0(x)p(x)\,\mathrm{d}x + \int \ell_0(x)\Delta_{\log \ell|s}(x)p(x)\,\mathrm{d}x$$

$$\Delta_{\log \ell|s} := m_{\log \ell|s} - \log \ell_0$$

Aprendizaje Activo de la Evidencia

- Como L_0 utiliza un GP estándar.
  - $\mathbf{L_0 = m_{L|s}}$
  - Prior con media 0
  - Covarianza Gaussiana
  - Utiliza el conjunto de datos X_S

- Utiliza un GP diferente para modelar $\Delta_{\log L|s}$
  - Utiliza prior con media 0
  - Covarianza Gaussiana
  - Para entrenar utiliza X_S y datos aleatorios en hyper-elipses alrededor de los puntos X_S

Aprendizaje Activo de Evidencia

- Quedando finalmente la media compuesta por dos términos analíticos.

$$m(Z|\log \ell_s) \simeq m(Z|\log \ell_0, \log \ell_s, \Delta_c) = m(Z|\ell_s) + m\big(\langle \ell \Delta_{\log \ell|s} \rangle \big| \ell_s, \Delta_c\big)$$

- Y una varianza que puede emplearse como diagnóstico de convergencia:

$$V(Z|\log \ell_0, \log \ell_s, \Delta_c) = S(Z\,|\log \ell_0, \log \ell_s) - m(Z|\log \ell_0, \log \ell_s, \Delta_c)^2$$

$$S(Z\,|\log \ell_0, \log \ell_s) := m\big(\langle \ell\, C_{\log \ell|s}\, \ell \rangle \big| \log \ell_s\big) + m(Z|\log \ell_0, \log \ell_s, \Delta_c)^2$$

$$V(Z|\log \ell_0, \log \ell_s, \Delta_c) = m\big(\langle \ell\, C_{\log \ell|s}\, \ell \rangle \big| \log \ell_s\big)$$
$$:= \iint m_{\ell|s}(x) m_{\ell|s}(x') C_{\log \ell|s}(x, x') p(x) p(x') \mathrm{d}x \mathrm{d}x'$$

Aprendizaje Activo de Evidencia

- Aprendizaje activo:
  - Ya no es necesario coger muestras que pertenezcan a p(x).
  - Cuando se han fijado los hyperparámetros, la varianza depende de la posición de las muestras escogidas.
  - Selecciona muestras que minimizan la varianza esperada:

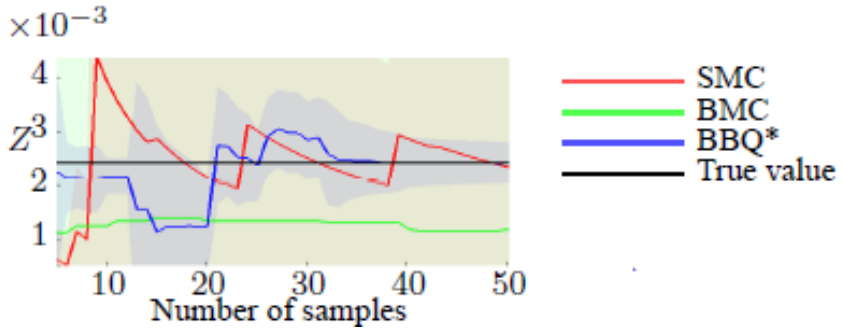
$$x_a = \operatorname{argmin}_{x_a} \left\langle V(Z|\log \ell_0, \log \ell_{s,a}) \mid \log \ell_0, \log \ell_s \right\rangle$$

$$\left\langle V(Z|\log \ell_0, \log \ell_{s,a}) \mid \log \ell_0, \log \ell_s \right\rangle = S(Z \mid \log \ell_0, \log \ell_s) - \int m(Z|\log \ell_0, \log \ell_{a,s}, \Delta_c)^2 \times \mathcal{N}\left(\log \ell_a; \hat{m}_a, \hat{C}_a + \frac{\partial \hat{m}_a}{\partial w} C_w \frac{\partial \hat{m}_a^{\mathsf{T}}}{\partial w} \right) \mathrm{dlog}\, \ell_a$$

$$\hat{m}_a := m(\log \ell_a | \log \ell_s, \hat{w})$$
$$\hat{C}_a := V(\log \ell_a | \log \ell_s, \hat{w})$$

Aprendizaje Activo de Evidencia

- Resultados:



$$\text{ALE} := \frac{1}{N}\sum_{i=1}^{N} |\log m(Z_i) - \log Z_i|$$

Synthetic Results

Method	ALE
SMC	1.101
AIS	1.695
BMC	2.695
BQ	6.760
BBQ	0.919

Real Results

Method	ALE
SMC	0.632
AIS	2.146
BMC	1.455
BQ	0.635
BBQ	0.400

Referencias

- www.probabilistic-numerics.org
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.
- Rasmussen, C. E., & Ghahramani, Z. Bayesian monte carlo. *Advances in neural information processing systems*, *15*, 489-496.
- Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S. J., & Ghahramani, Z. Active Learning of Model Evidence Using Bayesian Quadrature. NIPS 2012