## Motivación.

### Característ. ~~Ventajas~~ de Métodos Bayesianos.

- Accounting for uncertainty.

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

- Típicamente se requieren integrales, que hay que evaluar de forma aproximada.

  * Monte-Carlo sampling
  * Variational methods.

### Why be Bayesian with Big Data

- $p(\theta|\underline{x})$ se haría infinitamente estrecha en torno al máximo de $p(x|\theta)$, y tendríamos $\hat{\theta}_{MAP} = \hat{\theta}_{ML}$.
- Además $\int \hat{s}(\theta)\, p(\theta|\underline{x}^*)\, d\theta \longrightarrow \hat{s}(\hat{\theta}_{MAP})$
- En aplicaciones de big data, el nº de parámetros y la complejidad del modelo crece con el nº de datos.
- Retos:
  * Computacionales y de memoria: requerimos métodos secuenciales y paralelizables.
  * Nuevos métodos en los que puede ser deseable incurrir en mayores errores de aproximación, o permitir sesgo no nulo
  * Idealmente: Entender este tradeoff.

OUTLINE

0. MOTIVATION
1. EXPONENTIAL FAMILIES

THE EXPONENTIAL FAMILY OF DISTRIBUTIONS. (I)

- $X$ is a random vector, $X \in S_1 \subset \mathbb{R}^{d_1}$
- $\theta$ is a parameter vector, $\theta \in \Theta \subset \mathbb{R}^{d_2}$ (random)

DEF: $\{p(\cdot|\theta)\}$ es una familia exponencial, si cada densidad de la familia puede escribirse como:

$$p(x|\theta) = h(x)\, e^{\langle \eta(\theta), t(x) \rangle - \log Z(\eta(\theta))}$$

$$\longrightarrow \langle \eta(\theta), t(x) \rangle = [\eta(\theta)]^T t(x) = \sum_{i=1}^{K} \eta_i(\theta)\, t_i(x)$$

$\longrightarrow$ $\eta_i(\theta) \equiv$ PARÁMETROS NATURALES

$\longrightarrow$ $t_i(x) \equiv$ ESTADÍSTICOS NATURALES (SUFICIENTES) $\theta \perp\!\!\!\perp X \mid t(x)$

T.E.F.O.D. (II)

$\longrightarrow$ ASUMIMOS QUE $S_1$ NO DEPENDE DE $\theta$

$\longrightarrow$ LA FAMILIA ES REGULAR SI $S_2$ es un conjunto abierto.

$\longrightarrow$ LA FAMILIA ES MÍNIMA SI NO $\exists\, a \neq 0$ tal que $\langle a, t(x) \rangle = cte.$

$\longrightarrow$ OTRAS FORMAS FRECUENTES:

$$p(x|\theta) = \alpha(\theta)\, h(x) \exp\left[\langle \eta(\theta), t(x) \rangle\right]$$

$$p(x|\theta) = h(x) \exp\left[\langle \eta(\theta), t(x) \rangle - A(\theta)\right]$$

$$A(\theta) = \log\left[Z(\eta(\theta))\right]$$

## Examples

Gaussian

Bernoulli

Binomial

...

Ver tx. 3. de las que me descargué.

## NATURAL PARAMETER FORM FOR BERNOULLI:

$$p(x) = \theta^x (1-\theta)^{1-x}$$

$$= \exp\left[\log\left(\theta^x (1-\theta)^{1-x}\right)\right]$$

$$= \exp\left[x \log\theta + (1-x)\log(1-\theta)\right]$$

$$= \exp\left[x \log\frac{\theta}{1-\theta} + \log(1-\theta)\right]$$

Luego: $h(x) = 1$ ; $\eta(\theta) = \log\frac{\theta}{1-\theta}$

$t(x) = x$ ; $-\log Z(\eta) = \log(1-\theta)$

$$Z(\theta) = \frac{1}{1-\theta}$$

Además, como $\frac{\theta}{1-\theta} = e^\eta$ ; $\theta = e^\eta(1-\theta) = e^\eta - \theta e^\eta$

$$\theta = \frac{e^\eta}{1+e^\eta}$$

$$1-\theta = \frac{1}{1+e^\eta}$$

Luego $Z(\eta(\theta)) = 1 + e^\eta$

$$p(x|\theta) = \exp\left[x\,\eta - \log(1+e^\eta)\right]$$

## NATURAL PARAMETER FOR GAUSSIAN

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\log\sigma\right] \exp\left[\frac{-x^2 + 2\mu x - \mu^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right]$$

- $h(x) = \dfrac{1}{\sqrt{2\pi}}$
- $t(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$

- $\eta(\theta) = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix}$
- $Z(\theta) = \exp\left[\dfrac{\mu^2}{2\sigma^2} + \log\sigma\right]$

$$\eta_1 = \mu/\sigma^2 \quad \Rightarrow \quad \sigma^2 = \frac{-1}{2\eta_2}$$

$$\eta_2 = -1/2\sigma^2 \quad \Rightarrow \quad \mu = \eta_1 \cdot \sigma^2 = \frac{-\eta_1}{2\eta_2}$$

$$Z(\theta) = \exp\left[\frac{\eta_1^2}{4\eta_2^2} \cdot (-\eta_2) + \frac{1}{2}\log\frac{-1}{2\eta_2}\right]$$

$$= \exp\left[\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)\right]$$

## MUESTREO DE $p(x|\theta)$

Si la distribución de $X$ es de una familia exponencial con parámetros naturales $\eta_i(\theta)$ y estadísticos naturales $t_i(x)$, entonces una colección de $n$ muestras de $X = (X_1, \dots, X_n)$ también sigue una familia exponencial con los mismos parámetros naturales, y estadísticos naturales

$$t_i'(\underline{x}) = \sum_{k=1}^{n} t_i(x_k)$$

IMPT. FOR BIG DATA

Dem:

$$p(\underline{x}) = \prod_{k=1}^{n} p(x_k|\theta) = \prod_{k=1}^{n} h(x_k) \exp\left[\sum_i t_i(x_k)\eta_i(\theta) - \log Z(\theta)\right]$$

$$= \prod_{k=1}^{n} h(x_k) \exp\left[\sum_{k=1}^{n}\sum_i t_i(x_k)\eta_i(\theta) - \log Z(\theta)\right]$$

$$= \prod_{k=1}^{n} h(x_k) \exp\left[\sum_i \eta_i(\theta)\sum_{k=1}^{n} t_i(x_k) - \log Z(\theta)\right]$$

## THE ROLE OF $Z(\eta)$ / $A(\eta)$

- PERMITE NORMALIZAR LA DISTRIBUCIÓN, I.E.:

$$\int p(x|\theta)\,dx = \exp^{-A(\eta)} \int h(x) \exp\left[\eta^T t(x)\right] dx = 1$$

Luego $e^{A(\eta)} = Z(\eta) = \int h(x) \exp\left[\eta^T t(x)\right] dx$

i.e., si $t(x) = x$ tendríamos la T. Laplace.

- VAMOS A VER ALGUNAS VENTAJAS DE LA REPRESENTACIÓN NATURAL.

## DERIVADA DE $A(\eta)$

$$A(\eta) = \log Z(\eta) = \log \underbrace{\int h(x)\, e^{\eta^T \cdot t(x)}\, dx}_{Q(\eta)}$$

$$\frac{dA(\eta)}{d\eta} = \frac{1}{Q(\eta)} \cdot \frac{dQ(\eta)}{d\eta} = \frac{\int h(x)\, e^{\eta^T t} \cdot t(x)\, dx}{\int h(x)\, e^{\eta^T t}\, dx} \cdot \frac{e^{-A(\eta)}}{e^{-A(\eta)}}$$

$$= \frac{\int p(x) \cdot t(x)\, dx}{\int p(x)\, dx} = \mathbb{E}\left[t(x)\right]$$

E.g. Bernoulli:

$$A(\eta) = \log(1 + e^{\eta})$$

$$\frac{dA}{d\eta} = \frac{e^{\eta}}{1 + e^{\eta}} = \theta$$

Luego $\mathbb{E}\{t(x)\} = \mathbb{E}\{x\} = \theta$

E.g. Gaussian:

$$A(\eta) = \frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$$

$$\frac{dA}{d\eta_1} = \frac{-2\eta_1}{4\eta_2} = \frac{-\eta_1}{2\eta_2} = \mu \qquad \left[\mathbb{E}\{t_1(x)\} = \mathbb{E}\{x\}\right]$$

$$\frac{dA}{d\eta_2} = \frac{\eta_1^2}{4\eta_2^2} + \frac{1}{2}\frac{-2}{2\eta_2} = \mu^2 - \frac{1}{2\eta_2} = \mu^2 + \sigma^2 \left[\mathbb{E}\{t_2(x)\} = \mathbb{E}\{x^2\}\right]$$

## SEGUNDA DERIVADA DE $A(\eta)$

- De manera similar:

$$\nabla^2 A(\eta) = \mathbb{E}\left[t(x)\, t(x)^T\right] - \mathbb{E}\{t(x)\}\,\mathbb{E}\{t(x)\}^T$$

$$= \mathrm{Cov}\left[t(x)\right]$$

## MOMENT GENERATING FUNCION

- For a general r.v. $T$, its moment generating function is defined as $M_T(s) = \mathbb{E}\{e^{s \cdot T}\}$
- It does not always exist, but if it does it allows an easy calculation of moments, since:

$$\mathbb{E}\{e^{s \cdot T}\} = 1 + \mathbb{E}\{s \cdot T\} + \frac{\mathbb{E}\{(s \cdot T)^2\}}{2!} + \frac{\mathbb{E}\{(s \cdot T)^3\}}{3!} + \dots$$

Therefore $\nabla_{(s)}^k \mathbb{E}\{e^{s^T T}\}\Big|_{s=0} = \mathbb{E}\{T^k\}$

- FOR EXPONENTIAL FAMILIES:

$$\mathbb{E}\{e^{s^T t(x)}\} = \int e^{s^T t(x)} \cdot p(x)\, dx$$

$$= \int e^{s^T t(x)} \cdot h(x) \cdot e^{\eta^T t(x)}\, e^{-\log Z}\, dx$$

$$= e^{-\log Z} \int h(x)\, e^{(s+\eta)^T t}\, dx$$

$$= e^{-\log Z(\eta)} \cdot Z(s+\eta)$$

$$= e^{A(\eta+s) - A(\eta)}$$ Cumulant generating function.

$$\nabla_{(s)}^k \mathbb{E}\{e^{s^T t}\}\Big|_{s=0} = \nabla_{(s)}^k A(\eta+s)\Big|_{s=0} = \nabla_{(s)}^k A(\eta)$$

$$e^{(t+s)^3 - t^3}$$

$$\frac{d}{dt} = e^{(t+s)^3 - t^3}\Big|_{s=0} \cdot 3(t+s)^2 - 3t^2$$

## OTRAS PROPIEDADES

- Si la familia es regular, definimos el "score" con respecto al parámetro natural:

$$v(x;\eta) \triangleq \nabla_\eta \log p(x|\eta) = \nabla_\eta \left[ \langle \eta, t \rangle - \log Z(\eta) \right]$$

$$= t(x) - \nabla_\eta \log Z(\eta) = t(x) - \mathbb{E}\{t(x)\}$$

- Si la familia es regular, la información de Fisher con respecto al parámetro natural es:

$$I(\eta) \triangleq \mathbb{E}\{ v(x,\eta)\, v(x,\eta)^T \} = \mathbb{E}\{ t(x) - \mathbb{E}\{t(x)\} \}^2 = \mathrm{Cov}(t(x)) = \nabla^2 \log Z(\eta)$$

- SAMPLING
- ML solution

$$\nabla_\theta \log p(x) = \nabla_\theta \left[ \eta^T \sum_i t(x_i) - n \log Z(\eta) \right]$$

$A(\eta)$

luego en la solución se satisface

$$\nabla A(\hat{\eta}_{ML}) = \frac{1}{N} \sum_{i=1}^{N} t(x_i) = 0$$

# CONJUGATE PRIORS IN BAYESIAN STATISTICS.

$$p(\theta|x,\alpha') = \frac{p(\theta|\alpha)\cdot p(x|\theta)}{p(x)}$$

Si se cumple q. $p(\theta|\alpha)$ y $p(\theta|x,\alpha')$ son pdfs paramétricas (en $\alpha'$) de la misma familia, entonces decimos que el prior $p(\theta|\alpha)$ es conjugado con la función de verosimilitud $p(x|\theta)$

→ En general, $\alpha' = \alpha'(\alpha, x)$

→ La definición es general ...

→ pero si $p(x|\theta)$ pertenece a una familia exponencial regular, entonces siempre se puede encontrar un prior conjugado. (q. tb. es exponencial)

Def: • $p(x|\theta) = h_x(x)\exp\left[\langle \eta_x(\theta), t_x(x)\rangle - \log Z_x(\eta_x(\theta))\right]$

$= h_x(x)\exp\left[\langle (\eta_x(\theta), -\log Z_x(\,)), \underbrace{(t_x(x), 1)}_{t_\theta(\theta)}\rangle\right]$

• CONJ. PRIOR

$$p(\theta|\alpha) = h_\theta(\theta)\exp\left[\langle \eta_\theta(\alpha), \underbrace{(\eta_x(\theta), -\log Z_x(\eta(\theta)))}_{t_\theta(\theta)}\rangle - \log Z_\theta(\eta_\theta(\alpha))\right]$$

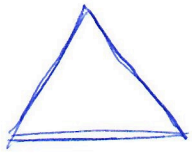
• POSTERIOR

$$p(\theta|x,\alpha) \propto h_\theta(\theta)\cdot\exp\left[\langle t_\theta(\theta), \eta_\theta(\alpha) + (t_x(x), 1)\rangle\right]$$

$t_\theta(\theta)$ → MISMOS ESTADÍSTICOS NATURALES

$\eta_\theta(\alpha) + (t_x(x), 1)$ → VALORES DIFERENTES DE LOS PARÁMETROS NATURALES

## Ejemplo: Dirichlet y Multinomial



- PRIOR: DIRICHLET. $p(\theta|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \cdot \prod_i \theta_i^{\alpha_i - 1}$

- MULTINOMIAL (VEROSIMILITUD): $p(x|\theta) = \frac{(\sum_i x_i)!}{x_1! x_2! \dots x_n!} \prod_{i=1}^{n} \theta_i^{x_i}$

  $x_1$ : # 1's
  $x_2$ : # 2's
  ⋮
  $x_n$ : # n's

  K experimentos
  n posibles valores.

- POSTERIOR: $p(\theta|x,\alpha) \propto \prod_{i=1}^{n} \theta_i^{(x_i+\alpha_i-1)}$

  q. es Dirichlet con parámetros $x_i + \alpha_i$

Por tanto, ha de ser:

$$p(\theta|x) = \frac{\Gamma(\sum_i \alpha_i + x_i)}{\prod_i \Gamma(\alpha_i + x_i)} \cdot \prod_{i=1}^{n} \theta_i^{(x_i+\alpha_i)-1}$$

Otros pares de conjugados:

- Gaussian – Gaussian.
- Beta – Bernoulli
- ...

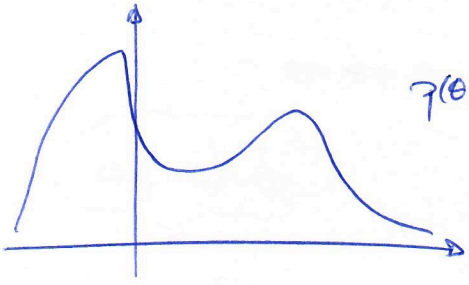
## MCMC INFERENCE

- Recordemos que en INF. BAYESIANA

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)}$$

- EN OCASIONES $\theta$, ADEMÁS DE HIPERPARÁMETROS TIENE VAR. LATENTES DEL MODELO

- OBJETIVO MCMC: obtener muestras i.i.d de $p(\theta|x)$ ~~CUANDO ESTA NO SE PUEDE EVALUAR DE FORMA CERRADA,~~ o CUANDO QUIERO EVALUAR INTEGRALES TIPO $\int_\theta f(\theta) \cdot p(\theta|x)\, d\theta$ (TB.)




~~ESTOS~~

MCMC: $\theta_0, \theta_1, \ldots, \theta_t, \theta_{t+1}$

BAJO DETERMINADAS CONDICIONES LA CADENA CONVERGE A LA DISTRIBUCIÓN BUSCADA, DE MANERA QUE PODEMOS QUEDARNOS CON UNA SERIE DE VALORES (NO CONSECUTIVOS).

- WE COLLECT ~~SAMPLES~~ FROM THE ~~CHAIN~~ SIMULATED TRAJECTORY AND USE THEM TO COMPUTE MONTE CARLO ESTIMATES.

## BIAS AND VARIANCE

- Imagine the estimation of r.variable $z$, using an estimator $\hat{z}$

  - $\text{BIAS}[\hat{z}] = \mathbb{E}\{\hat{z} - z\} = \mathbb{E}\{\hat{z}\} - \mathbb{E}\{z\}$
  - $\text{VAR}[\hat{z}] = \mathbb{E}\{(\hat{z} - \mathbb{E}\{\hat{z}\})^2\}$
  - $\text{MSE}[\hat{z}] = (\text{BIAS}[\hat{z}])^2 + \text{VAR}[\hat{z}]$

- TÍPICAMENTE CON MCMC SE DEJA ITERAR MUCHO Y $\text{BIAS}[\hat{z}] \to 0$; EN BIG DATA PUEDE SER PREFERIBLE NO ELIMINAR COMPLETAMENTE EL SESGO ASINTÓTICO

## MONTE CARLO ESTIMATOR

- We want to estimate:

$$\mathbb{E}[f(\theta)] = \int_{\Theta} f(\theta) \cdot p(\theta|x)\, d\theta$$

- Given a sequence of i.i.d samples $\{\theta_i\}$ from $p(\theta|x)$:

$$\hat{\mathbb{E}}[f(\theta)] = \frac{1}{n} \sum_{i=1}^{n} f(\theta_i)$$

- El CLT permite escribir que conforme $n \to \infty$

~~$\frac{1}{n} \sum_{i=1}^{n} f(\theta_i) - \mathbb{E}[f(\theta)] \sim \mathcal{N}(0, \sigma^2)$~~

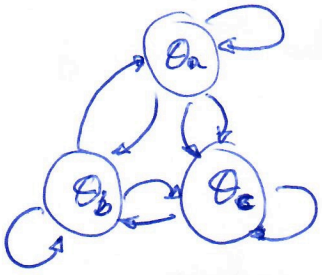
$\sigma^2 \sim \mathrm{Var}\, f(\theta)$

$$\frac{1}{n} \sum_{i=1}^{n} f(\theta_i) - \mathbb{E}[f(\theta)] \sim \mathcal{N}\left(0, (\sigma/\sqrt{n})^2\right)$$

- I.e., el error MCSE decrece con $\sigma$ escala con $1/\sqrt{n}$ independientemente de la dimensión de $\theta$.

# MARKOV CHAIN

## ESPACIO DE ESTADOS DISCRETO.




$\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \ldots \in \{\theta_a, \theta_b, \theta_c\}$

Distrib. inicial: $\pi_0(\theta)$

Prob. de transición: $P(\theta_{t+1} | \theta_t) : T(\theta_t \to \theta_{t+1})$

(PROP. MARKOV).

Homogéneo: Cuando las propiedades no dependen de $t$.

## ESPACIO DE ESTADOS ~~DISCRETO~~ CONTINUO

$\theta_0, \theta_1, \theta_2, \theta_3, \ldots, \theta_t, \ldots \in \mathbb{R}$

HOMOGÉNEO EN EL TIEMPO. Prob. de transición: $T(\theta \to \theta')$

- $\pi_t(\theta)$ es la d.d.p. de $\theta_t$, depende de $T(\theta \to \theta')$, pero también de $\pi_0(\theta)$

- $\pi_{t+1}(\theta) = (\pi T)(\theta) = \int_{\mathbb{R}} T(\theta' \to \theta)\, \pi_t(\theta')\, d\theta'$

# MARKOV CHAIN: SIMULATION

- Sample $\theta_0$ from $\pi_0(\theta)$
- For $t = 0, 1, 2, \ldots$

  Sample $\theta_{t+1}$ from $p(\theta_{t+1} | \theta_t) : T(\theta_t \to \theta)$

- Nos interesan las cadenas q. convergen a una distribución estacionaria única

$$\lim_{t \to \infty} \pi_t(\theta) = \pi(\theta)$$

independientemente de la distribución $\pi_0(\theta)$

- <u>Objetivo</u>: Diseñar $T$ para q. $\pi(\theta)$ sea la deseada.

## MARKOV CHAIN: DISTRIB. ESTACIONARIA

- $\pi(\theta)$ es una distrib. estacionaria para un operador $T$, si y solo si

$$\pi(\theta) = (\pi T)\,\theta$$

- Existen condiciones para el estudio de cuando una cadena converge a una distrib. estacionaria única.

- Condición suficiente:

$$\underbrace{T(\theta \to \theta')\,\pi(\theta)}_{\text{Prob. q. mueve a } \theta'} = \underbrace{T(\theta' \to \theta)\,\pi(\theta')}_{\text{Prob. que sale de } \theta'} \quad \forall \theta, \theta'$$

- Decimos q. se satisface "the detailed balance condition" o que el operador $T$ es reversible.

## MCMC ESTIMATION

- Sea $\pi(x)$ la distrib. estacionaria de la MC, podemos estimar:

$$\mathbb{E}[f(\theta)] = \int_{\Theta} f(\theta)\,\pi(\theta)\,d\theta \approx \frac{1}{n}\sum_{i=1}^{n} f(\theta_i)$$

- Nótese q. no hemos exigido i.i.d., aún así, se satisfacen resultados asintóticos similares a los del est. de Montecarlo.

$$\frac{1}{n}\sum_{i=1}^{n} f(\theta_i) - \mathbb{E}\{f(\theta)\} \sim \mathcal{N}\left(0, (\sigma/\sqrt{n})^2\right)$$

$$\text{con } \sigma^2 = \mathrm{Var}[f(\theta)] + 2\underbrace{\sum_{t=1}^{\infty} \mathrm{Cov}[f(\theta_0), f(\theta_t)]}_{\text{Penaliz. pq. las muestras no son i.i.d.}} \quad \left(\text{Tiende a } 0 \text{ conforme } t \to \infty\right)$$

(en la práctica es algo peor, pq. la expresión asume inicialización perfecta $\theta_0 \sim \pi(\theta)$ )

- En cualquier caso, decrece con $1/\sqrt{n}$ el error.

## MCMC ≡ Transient bias

- Previous results hold for the stationary phase of the Markov Chain.
- There is a transient bias due to initializing the Markov chain out of stationarity.

(If it were possible to initialize in the stationary case, then we could draw i.i.d. samples).

- Transient bias decreases at a rate at least $\mathcal{O}(1/n)$, and hence it is dominated by the MC standard error.

Fig. 2.1.

## MCMC ≡ Transient bias (2)

- It would be possible to remove transient bias. For instance, for very large $n$: $\hat{\mathbb{E}}\{f(\theta)\} = f(X_n)$

  ... but error would be fixed.

- More clever choices: $\hat{\mathbb{E}}\{f(\theta)\} = \frac{1}{\lceil n/2 \rceil} \sum_{i=\lceil n/2 \rceil}^{n} f(X_i)$

Fig. 2.1.

Fig. 2.2.

## METROPOLIS HASTINGS.

ALG. 1.

- $p(\theta, x)$ can be evaluated up to proportionality
- $q(\theta'|\theta) \sim N(\theta, \sigma)$, $\sigma$ ej.; tendrá influencia en el comportamiento del algoritmo.
- If $q$ symmetric simplifies to the Hastings algorithm.

## M.-H.- proof.

$$T(\theta' \to \theta)\, p(\theta'|x) =$$

$$= p(\theta'|x) \cdot q(\theta|\theta') \cdot \min\left[1, \frac{p(\theta|x)\, q(\theta'|\theta)}{p(\theta'|x)\, q(\theta|\theta')}\right]$$

$$= \min\left[p(\theta'|x) \cdot q(\theta|\theta'),\ p(\theta|x)\, q(\theta'|\theta)\right]$$

$$= \min\left[1, \frac{p(\theta'|x)\, q(\theta|\theta')}{p(\theta|x) \cdot q(\theta'|\theta)}\right] \cdot p(\theta|x) \cdot q(\theta'|\theta)$$

$$= T(\theta \to \theta') \cdot p(\theta|x)$$

c.q.d.

## GIBBS SAMPLING.

$$p(\theta|x) = p(\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(d)}|x)$$

Alg.: Initialize $\theta_0$

For $t = 0, \dots, n$:

For $j = 0, \dots, d$

Sample $\theta_{t+1}^{(j)}$ from $p(\theta^{(j)}|\theta_{t+1}^{(0)} \dots \theta_{t+1}^{(j-1)}\, \theta_t^{(j+1)} \dots \theta_t^{(n)}, x)$

a conjunta.

## STOCHASTIC GRADIENT ASCENT.

- OBJ: $\phi^* = \arg\max_{\phi} f(\phi)$

- $f(\phi) = \sum_{k=1}^{K} g(\phi, \bar{y}^{(k)})$ ; donde $\bar{y}^{(k)}$ es un minibatch de $k$ tos.

- $\nabla f(\phi) = \sum_{k=1}^{K} P_k \frac{1}{P_k} \nabla_\phi g(\phi, \bar{y}^{(k)}) = \mathbb{E}_{\hat{k}} \left[ \frac{1}{P_{\hat{k}}} \nabla_\phi g(\phi, \bar{y}^{(\hat{k})}) \right]$

ALG: Initialize $\phi^{(0)} \in \mathbb{R}^n$

for $t = 0, 1, 2, \ldots$

$\hat{k}_t \leftarrow$ sample index $k$ with probability $P_k$ ; $k = 1, \ldots, K$

$\phi^{(t+1)} = \phi^{(t)} + \frac{\rho^{(t)}}{P_k} G^{(t)} \nabla_\phi g(\phi^{(t)}, \bar{y}^{(\hat{k}^{(t)})})$

## COND. TH. 2.11

- No general theory to analyze rates of convergence for non-convex problems.