

Inference and Representation

Lecture 7

Joan Bruna
Courant Institute, NYU



Lecture 7 Objectives

- Markov-Chain Monte-Carlo (MCMC)
- Interlude: The Stein Operator
- Variational Inference

Markov Chains flash review

- We consider:

A state-space Ω (discrete or continuous)

A *stochastic* kernel P that updates the densities at each time step:

$$q^{(t+1)}(u) = \int_{\Omega} P(u, v)q^{(t)}(v)dv .$$

$$P(u, v) \geq 0 , \quad \int P(u, v)dv = 1 , \quad \forall u \in \Omega .$$

Markov Chains flash review

- We consider:
 - A state-space Ω (discrete or continuous)
 - A *stochastic* kernel P that updates the densities at each time step:

$$q^{(t+1)}(u) = \int_{\Omega} P(u, v) q^{(t)}(v) dv .$$
$$P(u, v) \geq 0 , \quad \int P(u, v) dv = 1 , \quad \forall u \in \Omega .$$

called
stochastic
matrix

- In a discrete state space, the kernel can be viewed as a matrix, and

$P(k, l)$ = Probability to transition from state k to state l .

Markov-Chains Flash Review

Definition: A Markov Chain is irreducible if
 $Pr(\text{visit state } u \text{ from state } v) > 0$ for all $u, v \in \Omega$.

Definition: A Markov Chain is aperiodic if there exists $n > 0$ such that
 $Pr(X_{n'} = i \mid X_0 = i) > 0$ for all $n' \geq n$. $\forall i$

Definition: A stationary distribution π of a Markov Chain P is such that $\pi = P\pi$.

Markov-Chains Flash Review

Definition: A Markov Chain is irreducible if $\Pr(\text{visit state } u \text{ from state } v) > 0$ for all $u, v \in \Omega$.

Definition: A Markov Chain is aperiodic if there exists $n > 0$ such that $\Pr(X_{n'} = i \mid X_0 = i) > 0$ for all $n' \geq n$. $\forall i$

Definition: A stationary distribution π of a Markov Chain P is such that $\pi = P\pi$.

Theorem: [Perron-Frobenius] An irreducible aperiodic Markov Chain P admits a unique stationary distribution.

Remark: The stationary distribution(s) correspond to eigenvectors of P of leading eigenvalue ($= 1$).

Metropolis-Hastings Algorithm

Definition: A Markov Chain P is reversible with respect to π if

$$\forall u, v \in \Omega, \pi(u)P(v, u) = \pi(v)P(u, v).$$

Fact: If P is reversible wrt π then π is a stationary for P .

Metropolis-Hastings Algorithm

Definition: A Markov Chain P is reversible with respect to π if

$$\forall u, v \in \Omega, \pi(u)P(v, u) = \pi(v)P(u, v).$$

Fact: If P is reversible wrt π then π is a stationary for P .

$$(P\pi)(u) = \sum_v P(u, v)\pi(v) = \sum_v \pi(u)P(v, u) = \pi(u)$$

- How to build a Markov Chain such that it is reversible wrt $p(x)$? without involving the partition function?

Metropolis-Hastings

- We start with a *proposal* Markov chain K such that
 - $\forall i \in \Omega, K_{i,i} > 0$, and
 - $G = (\Omega, E(K))$ is connected, where i and j are connected if $K_{i,j}K_{j,i} > 0$.
- (we can travel from any state to any other state in finite time using K).

Metropolis-Hastings

- We start with a *proposal* Markov chain K such that
 - $\forall i \in \Omega, K_{i,i} > 0$, and
 - $G = (\Omega, E(K))$ is connected, where i and j are connected if $K_{i,j}K_{j,i} > 0$.

(we can travel from any state to any other state in finite time using K).

- Then we adapt it to $p(x)$ as follows:

$$R(x_j, x_i) = \min \left\{ 1, \frac{\tilde{p}(x_j)K(x_j, x_i)}{\tilde{p}(x_i)K(x_i, x_j)} \right\},$$

$$P(x_j, x_i) = K(x_j, x_i)R(x_j, x_i) + \delta(x_j - x_i)r(x_i), \text{ with}$$

$$r(x_i) = \int_{\Omega} K(y, x_i)(1 - R(y, x_i))dy.$$

= Probability to reject the update.

Metropolis-Hastings

Fact: The kernel P is irreducible, aperiodic and satisfies the detailed balance condition wrt p .

Therefore p is the unique stationary distribution of P .

diagonal part ($\delta(x_i - x_j)r(x_i)$) can be omitted (why?)

Assume wlog $\tilde{p}(x_j)K(x_j, x_i) \geq \tilde{p}(x_i)K(x_i, x_j)$.

Then $R(x_j, x_i) = 1$ and

$$R(x_i, x_j) = \frac{\tilde{p}(x_i)K(x_i, x_j)}{\tilde{p}(x_j)K(x_j, x_i)} = \frac{p(x_i)K(x_i, x_j)}{p(x_j)K(x_j, x_i)}.$$

$$p(x_i)P(x_j, x_i) = p(x_i)K(x_j, x_i) = p(x_i)K(x_j, x_i) \frac{p(x_j)K(x_i, x_j)}{p(x_j)K(x_i, x_j)}$$

$$= \frac{p(x_i)K(x_j, x_i)}{p(x_j)K(x_i, x_j)} p(x_j)K(x_i, x_j)$$

$$= R(x_i, x_j)K(x_i, x_j)p(x_j) = p(x_j)P(x_i, x_j).$$

Irreducible by assumption on K .

(ex: Why is it aperiodic?)

Metropolis-Hastings

- Practical Implementation:

1. Given current state \boldsymbol{x}_i , propose next step by sampling $\boldsymbol{x}_s \sim K(\cdot, \boldsymbol{x}_i)$
2. Compute
$$R = \min \left(1, \frac{p(\boldsymbol{x}_s)K(\boldsymbol{x}_s, \boldsymbol{x}_i)}{p(\boldsymbol{x}_i)K(\boldsymbol{x}_i, \boldsymbol{x}_s)} \right).$$
3. Accept new state \boldsymbol{x}_s with probability R and go to step 1.

Metropolis-Hastings

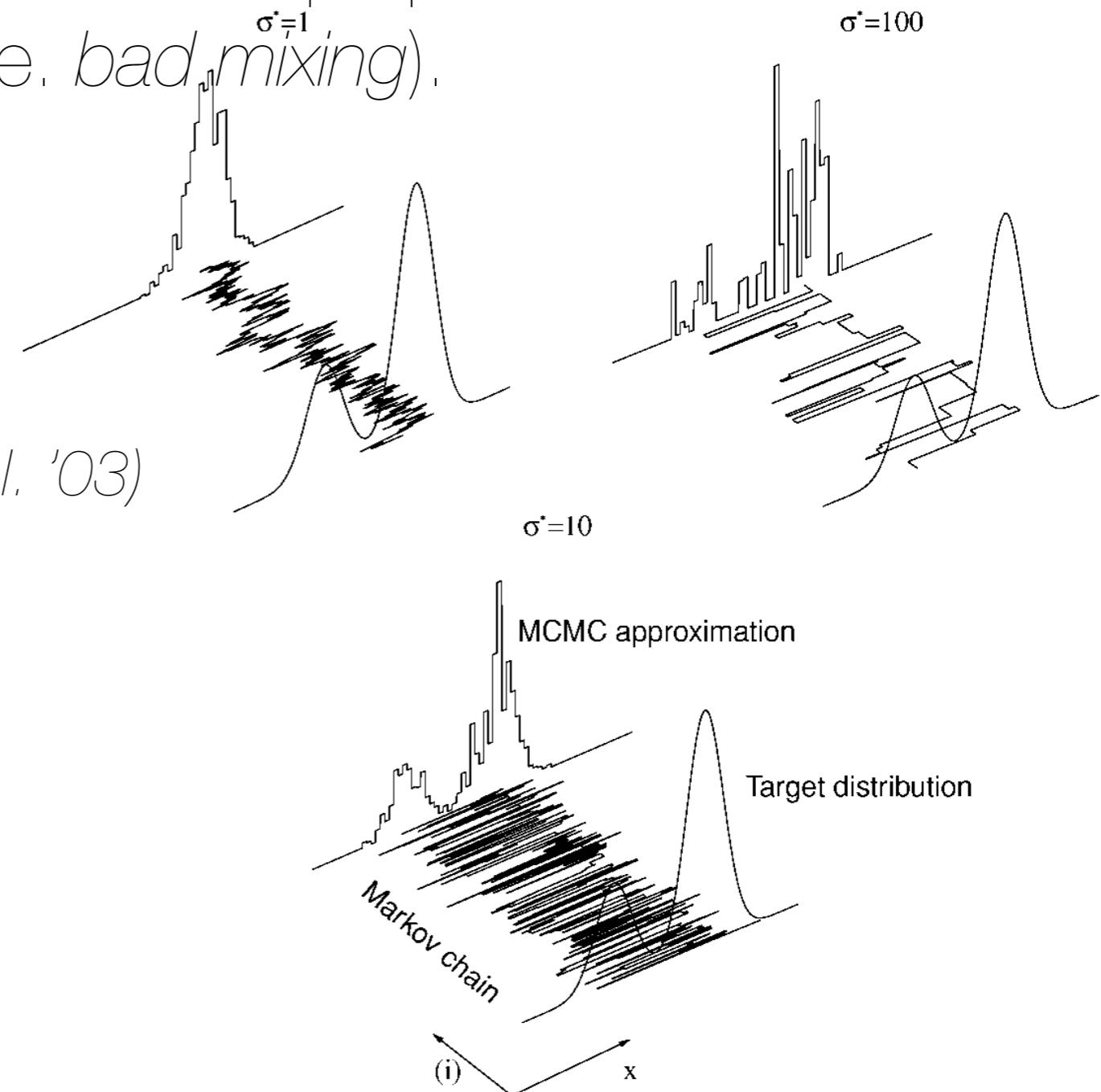
- Practical Implementation:
 1. Given current state x_i , propose next step by sampling $x_s \sim K(\cdot, x_i)$
 2. Compute $R = \min \left(1, \frac{p(x_s)K(x_s, x_i)}{p(x_i)K(x_i, x_s)} \right)$.
 3. Accept new state x_s with probability R and go to step 1.

- Two important particular cases:
 - $K(x_j, x_i) = q(x_j)$ is independent of the current state.
Acceptance probability becomes
$$R(x_i, x_j) = \min \left(1, \frac{p(x_j)q(x_i)}{p(x_i)q(x_j)} \right) = \min \left(1, \frac{w(x_j)}{w(x_i)} \right).$$
 - Similar to importance sampling, except that here samples are correlated (why?)
 - $K(x_j, x_i) = K(x_i, x_j)$ is symmetric.
Acceptance probability becomes $R(x_i, x_j) = \min \left(1, \frac{p(x_j)}{p(x_i)} \right)$.
 - We always accept the sample as long as it “climbs” wrt the true density.

Metropolis-Hastings

- This algorithm seems to be *magic*: it only asks us to cover the target density with our proposal distribution!
- There is no “free lunch”: a poor choice of proposal distribution results in bad approximations (i.e. *bad mixing*).

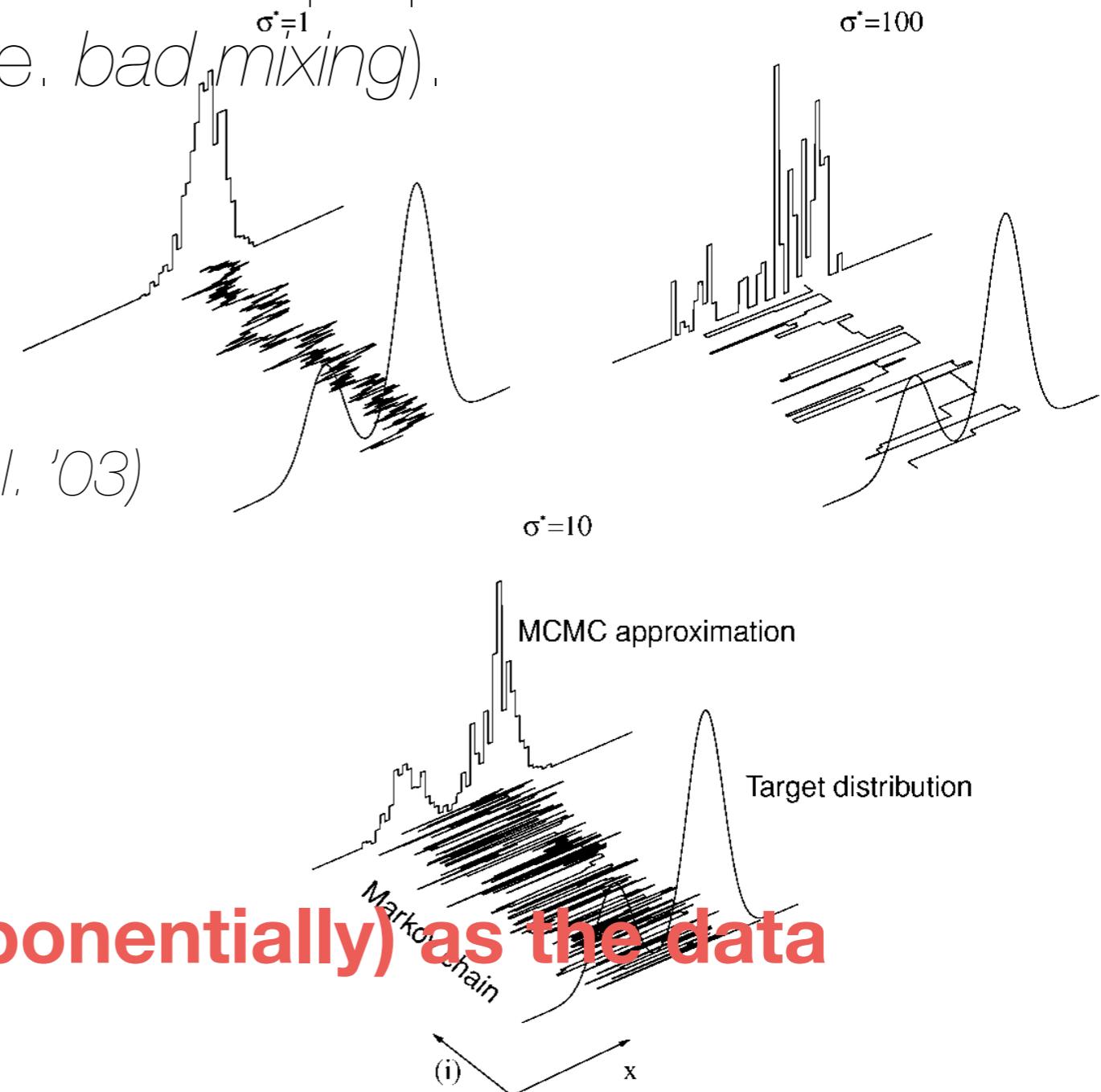
(figure from Andrieu et al. '03)



Metropolis-Hastings

- This algorithm seems to be *magic*: it only asks us to cover the target density with our proposal distribution!
- There is no “free lunch”: a poor choice of proposal distribution results in bad approximations (i.e. *bad mixing*).

(figure from Andrieu et al. '03)



- **This danger increases (exponentially) as the data dimensionality increases.**

Revisiting Gibbs Sampling

- Recall the Gibbs sampling we saw in Lecture 3.
 1. Select $k \in \{1, \dots, n\}$ from a uniform distribution.
 2. Set $x_j^{(n+1)} = x_j^{(n)}$ for $j \neq k$.
 3. Sample $x_k^{(n+1)}$ from $p(x_k \mid x_{-k}^{(n)})$.
- Q: Is this a “correct” algorithm? i.e, can we guarantee that samples from this algorithm come from $p(x)$?
 - Consider $K(y, x) = \begin{cases} p(y_k \mid x_{-k}) & \text{if } y_{-k} = x_{-k} \\ 0 & \text{otherwise.} \end{cases}$If $p(\cdot \mid x_{-k}) > 0$ then the resulting graph for K is connected
⇒ its resulting Markov Chain is irreducible and aperiodic.
- Does it satisfy detailed balance wrt $p(x)$?

Revisiting Gibbs Sampling

Lemma: The kernel K satisfies detailed balance wrt $p(x)$.

Revisiting Gibbs Sampling

Lemma: The kernel K satisfies detailed balance wrt $p(x)$.

We need to show that $p(x)K(x', x) = p(x')K(x, x')$.

Suppose that $x \neq x'$ and they differ in exactly one position k .

$$\begin{aligned} p(x)K(x', x) &= n^{-1}p(x)p(x'_k | x_{-k}) \\ &= n^{-1}p(x_k | x_{-k})p(x_{-k})p(x'_k | x_{-k}) \\ &= n^{-1}p(x_k | x'_{-k})p(x'_{-k})p(x'_k | x'_{-k}) \\ &= p(x')K(x, x') . \end{aligned}$$

Therefore, here $K = P$ (we accept the sample with probability 1).

Mixing Time in MCMC

- We have seen that the correctness of MCMC relies on a Markov chain having the appropriate stationary distribution.

Mixing Time in MCMC

- We have seen that the correctness of MCMC relies on a Markov chain having the appropriate stationary distribution.
- We have also seen that this stationary distribution is associated with the leading eigenvector of its kernel.
i.e, given initial guess μ , $\mu P^t \rightarrow \pi$ as $t \rightarrow \infty$.

Mixing Time in MCMC

- We have seen that the correctness of MCMC relies on a Markov chain having the appropriate stationary distribution.
- We have also seen that this stationary distribution is associated with the leading eigenvector of its kernel.
 - i.e, given initial guess μ , $\mu P^t \rightarrow \pi$ as $t \rightarrow \infty$.
- How fast does the chain converge to its leading eigenvector?
 - i.e. how small $\|\mu P^t - \pi\|$ is as $t \rightarrow \infty$?

Mixing Time in MCMC

- We have seen that the correctness of MCMC relies on a Markov chain having the appropriate stationary distribution.
- We have also seen that this stationary distribution is associated with the leading eigenvector of its kernel.
 - i.e, given initial guess μ , $\mu P^t \rightarrow \pi$ as $t \rightarrow \infty$.
- How fast does the chain converge to its leading eigenvector?
 - i.e. how small $\|\mu P^t - \pi\|$ is as $t \rightarrow \infty$?
- The convergence is dictated by the second *largest* eigenvalue:

$$\|\mu P^t - \pi\|_{TV} \leq \lambda_2^t \sqrt{\frac{1}{\min_j \pi_j}} .$$

Mixing Time in MCMC

- We have seen that the correctness of MCMC relies on a Markov chain having the appropriate stationary distribution.
- We have also seen that this stationary distribution is associated with the leading eigenvector of its kernel.

i.e, given initial guess μ , $\mu P^t \rightarrow \pi$ as $t \rightarrow \infty$.

- How fast does the chain converge to its leading eigenvector?
i.e. how small $\|\mu P^t - \pi\|$ is as $t \rightarrow \infty$?
- The convergence is dictated by the second *largest* eigenvalue:

$$\|\mu P^t - \pi\|_{TV} \leq \lambda_2^t \sqrt{\frac{1}{\min_j \pi_j}} .$$

– Second largest eigenvalues can be bounded using the Cheeger bound:

$$\lambda_2 \leq 1 - \frac{\Phi^2}{2} , \text{ where}$$

Φ is the conductance of P :

$$\min_{\Omega' \subset \Omega} \frac{\sum_{i \in \Omega', j \notin \Omega'} \pi_i P(i, j)}{\pi(\Omega')(1 - \pi(\Omega'))} .$$

Interlude: the Stein Method

- More generally, how can we evaluate whether a sample $\{x_1, \dots, x_n\}$ is legitimate for a certain $p(x)$, possibly known up to normalization?

Interlude: the Stein Method

- More generally, how can we evaluate whether a sample $\{x_1, \dots, x_n\}$ is legitimate for a certain $p(x)$, possibly known up to normalization?
- The *Stein method* (70') is a powerful tool to control the distance between two probability distributions using a metric of the form

$$d(P, Q) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{X \sim P} h(X) - \mathbb{E}_{Y \sim Q} h(Y)| ,$$

for a given family of functions $\mathcal{H} = \{h : \Omega \rightarrow \mathbb{R}\}$.

Interlude: the Stein Method

- More generally, how can we evaluate whether a sample $\{x_1, \dots, x_n\}$ is legitimate for a certain $p(x)$, possibly known up to normalization?
- The *Stein method* (70') is a powerful tool to control the distance between two probability distributions using a metric of the form

$$d(P, Q) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{X \sim P} h(X) - \mathbb{E}_{Y \sim Q} h(Y)| ,$$

for a given family of functions $\mathcal{H} = \{h : \Omega \rightarrow \mathbb{R}\}$.

- If we know Q , we know how its moments $E_{Y \sim f(Q)}$ behave for different f .

Stein's method reverses this: can we characterize the distribution by finding functional relationships between moments?

Interlude: the Stein Method

Suppose that Q is fixed, known distribution.

We look for an operator \mathcal{A} acting on functions $f : \Omega \rightarrow \mathbb{R}$ such that

$$\mathbb{E}_Y\{(\mathcal{A}f)Y\} = 0 \text{ for all } f \in \mathcal{F} \Leftrightarrow Y \text{ has distribution } Q.$$

Interlude: the Stein Method

Suppose that Q is fixed, known distribution.

We look for an operator \mathcal{A} acting on functions $f : \Omega \rightarrow \mathbb{R}$ such that

$$\mathbb{E}_Y\{(\mathcal{A}f)Y\} = 0 \text{ for all } f \in \mathcal{F} \Leftrightarrow Y \text{ has distribution } Q.$$

Ex: If Q is a standard normal $\mathcal{N}(0, 1)$, we have

$$\mathbb{E}\{f'(Y) - Yf(Y)\} = 0 \text{ for all } f \in C^1 \Leftrightarrow Y \sim \mathcal{N}(0, 1) .$$

(Stein's lemma)

So we can consider $(\mathcal{A}f)(x) = f'(x) - xf(x)$ in that case.

Interlude: the Stein Method

Suppose that Q is fixed, known distribution.

We look for an operator \mathcal{A} acting on functions $f : \Omega \rightarrow \mathbb{R}$ such that

$$\mathbb{E}_Y\{(\mathcal{A}f)Y\} = 0 \text{ for all } f \in \mathcal{F} \Leftrightarrow Y \text{ has distribution } Q.$$

Ex: If Q is a standard normal $\mathcal{N}(0, 1)$, we have

$$\mathbb{E}\{f'(Y) - Yf(Y)\} = 0 \text{ for all } f \in C^1 \Leftrightarrow Y \sim \mathcal{N}(0, 1) .$$

(Stein's lemma)

So we can consider $(\mathcal{A}f)(x) = f'(x) - xf(x)$ in that case.

Thus, if $P = Q$, then $\mathbb{E}_{X \sim P}\{(\mathcal{A}f)(X)\} = 0$ for all $f \in C^1$.

Hope: if $P \approx Q$, then $\mathbb{E}_{X \sim P}\{(\mathcal{A}f)(X)\} \approx 0$ for all $f \in C^1$.

Interlude: the Stein Method

Given a test function h , in some cases we can find $f = f_h$ such that

$$(\mathcal{A}f)(x) = h(x) - \mathbb{E}_{Y \sim Q}(h(Y)) .$$

Interlude: the Stein Method

Given a test function h , in some cases we can find $f = f_h$ such that

$$(\mathcal{A}f)(x) = h(x) - \mathbb{E}_{Y \sim Q}(h(Y)) .$$

It results that

$$\mathbb{E}_{W \sim P}\{(\mathcal{A}f)W\} = \mathbb{E}_{W \sim P}\{h(W)\} - \mathbb{E}_{Y \sim Q}\{h(Y)\} .$$

So we can control $\text{dist}(P, Q)$ by bounding the functional form on the lhs.

When is bounding lhs easier than bounding rhs?

Interlude: the Stein Method

Given a test function h , in some cases we can find $f = f_h$ such that

$$(\mathcal{A}f)(x) = h(x) - \mathbb{E}_{Y \sim Q}(h(Y)) .$$

It results that

$$\mathbb{E}_{W \sim P}\{(\mathcal{A}f)W\} = \mathbb{E}_{W \sim P}\{h(W)\} - \mathbb{E}_{Y \sim Q}\{h(Y)\} .$$

So we can control $\text{dist}(P, Q)$ by bounding the functional form on the lhs.

When is bounding lhs easier than bounding rhs?

Two very important special cases:

- Gaussian Case: $(\mathcal{A}f)(x) = f'(x) - xf(x)$, with
$$f(x) = e^{x^2/2} \int_{-\infty}^x [h(s) - \mathbb{E}(h(Y))]e^{-s^2/2} ds.$$
- Markov case: $(\mathcal{A}f)(x) = \langle f(x), \nabla \log p(x) \rangle + (\text{div } f)(x)$.

Crucial point: The Stein operator does not depend upon normalizing constant Z !

Evaluating sample quality

- Gorham and Mackey (NIPS'15) exploit the Stein operator to evaluate samples out of any MCMC algorithm.
 - Algorithm solves a linear program that bounds

$$\sup_{f \in \mathcal{H}} \sum_{i \leq n} (\langle f(x_i), \nabla \log p(x_i) \rangle + \operatorname{div} f(x_i)) .$$

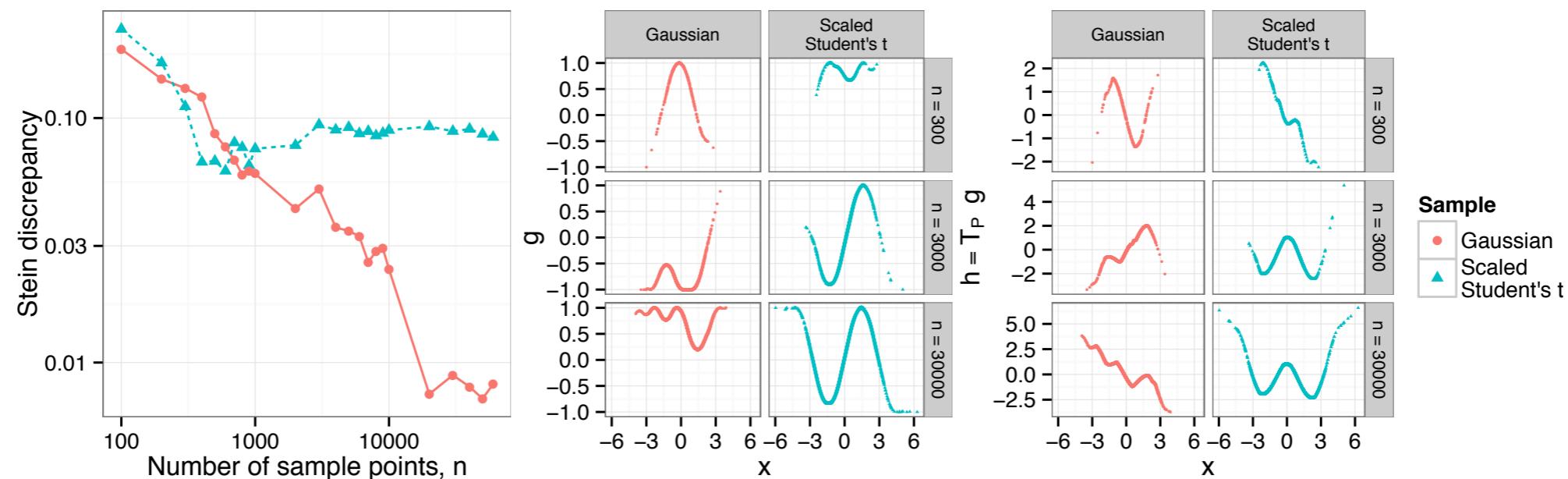


Figure 1: Left: Complete graph Stein discrepancy for a $\mathcal{N}(0, 1)$ target. Middle / right: Optimal Stein functions g and discriminating test functions $h = \mathcal{T}_P g$ recovered by the Stein program.

Evaluating sample quality

- Gorham and Mackey (NIPS'15) exploit the Stein operator to evaluate samples out of any MCMC algorithm.
 - Algorithm solves a linear program that bounds

$$\sup_{f \in \mathcal{H}} \sum_{i \leq n} (\langle f(x_i), \nabla \log p(x_i) \rangle + \operatorname{div} f(x_i)) .$$

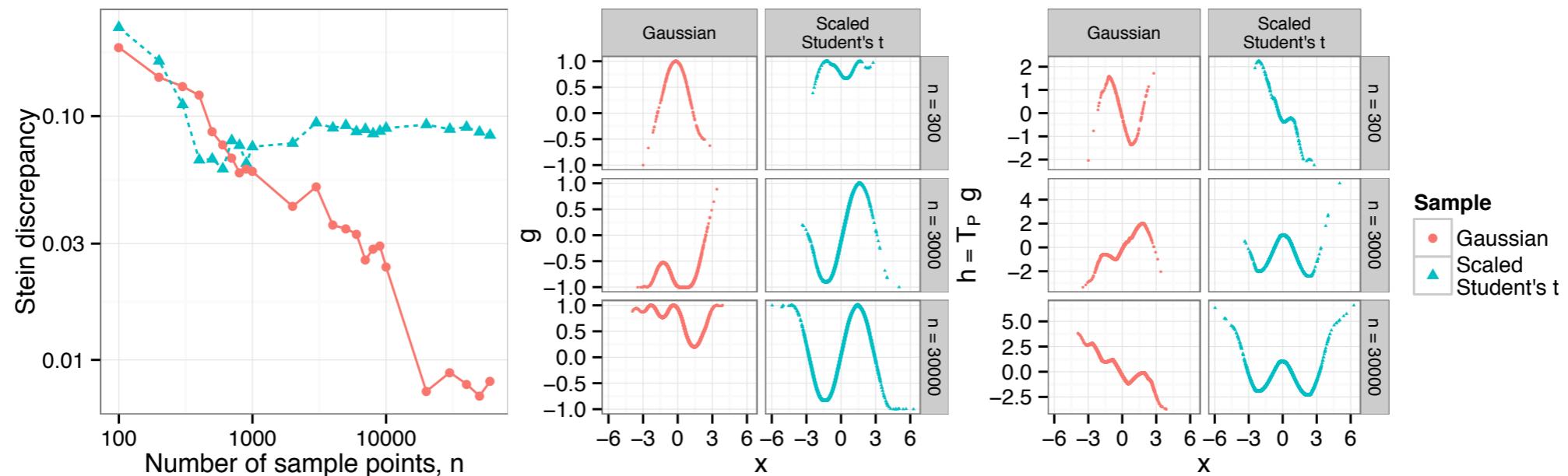


Figure 1: Left: Complete graph Stein discrepancy for a $\mathcal{N}(0, 1)$ target. Middle / right: Optimal Stein functions g and discriminating test functions $h = T_P g$ recovered by the Stein program.

- No assumption on the origin of the samples.
- Similar to Maximum-Mean Discrepancy (MMD) but Stein's method operates under weaker assumptions and is more efficient.

MCMC Extensions

- Hamiltonian Monte-Carlo (HMC) [Duane et al.'87][Neal'11]

Embeds a Gibbs distribution of the form $p(x) = \frac{e^{-T^{-1}U(x)}}{Z}$ into

$$p(x, y) = \frac{e^{-T^{-1}(U(x)+K(y))}}{Z}, \quad \begin{array}{l} x: \text{position} \\ y: \text{momentum} \end{array}$$

$U(x)$: potential energy $K(y)$: kinetic energy

$K(y) \propto \|y\|^2 \Rightarrow p(y|x)$ is Gaussian .

MCMC Extensions

- Hamiltonian Monte-Carlo (HMC) [Duane et al.'87][Neal'11]

Embeds a Gibbs distribution of the form $p(x) = \frac{e^{-T^{-1}U(x)}}{Z}$ into

$$p(x, y) = \frac{e^{-T^{-1}(U(x)+K(y))}}{Z}, \quad \begin{array}{l} x: \text{position} \\ y: \text{momentum} \end{array}$$

$U(x)$: potential energy $K(y)$: kinetic energy

$K(y) \propto \|y\|^2 \Rightarrow p(y|x)$ is Gaussian .

- Langevin Dynamics. [see Neal'10]
Discretization of a stochastic differential equation whose equilibrium distribution is the posterior

$$\nabla \theta_t = \frac{\epsilon}{2} \left(\nabla \log p(\theta_t) + \sum_{i \leq N} \nabla \log p(x_i \mid \theta_t) \right) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \epsilon).$$

MCMC Extensions

see ps5

- Hamiltonian Monte-Carlo (HMC) [Duane et al.'87][Neal'10]
Embeds a Gibbs distribution of the form $p(x) = \frac{e^{-T^{-1}U(x)}}{Z}$ into

$$p(x, y) = \frac{e^{-T^{-1}(U(x)+K(y))}}{Z}, \quad \begin{array}{l} x: \text{position} \\ y: \text{momentum} \end{array}$$

$U(x)$: potential energy $K(y)$: kinetic energy

$K(y) \propto \|y\|^2 \Rightarrow p(y|x)$ is Gaussian .

- Langevin Dynamics. [see Neal'10]
Discretization of a stochastic differential equation whose equilibrium distribution is the posterior

$$\nabla \theta_t = \frac{\epsilon}{2} \left(\nabla \log p(\theta_t) + \sum_{i \leq N} \nabla \log p(x_i \mid \theta_t) \right) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \epsilon).$$

Variational Inference

- Q: What does *variational* mean?

Variational Inference

- Q: What does *variational* mean?
- In general, it refers to the idea of expressing a quantity of interest θ^* (e.g. a posterior probability) as the solution of an optimization problem:

$$\theta^* = \inf_{\theta \in \mathcal{M}} f(\theta) .$$

Variational Inference

- Q: What does *variational* mean?
- In general, it refers to the idea of expressing a quantity of interest θ^* (e.g. a posterior probability) as the solution of an optimization problem:

$$\theta^* = \inf_{\theta \in \mathcal{M}} f(\theta) .$$

- Approximating the solution can now be accomplished by
 - Simplifying the domain \mathcal{M} .
 - Simplifying the function f .

Variational Inference

- Q: What does *variational* mean?
- In general, it refers to the idea of expressing a quantity of interest θ^* (e.g. a posterior probability) as the solution of an optimization problem:

$$\theta^* = \inf_{\theta \in \mathcal{M}} f(\theta) .$$

- Approximating the solution can now be accomplished by
 - Simplifying the domain \mathcal{M} .
 - Simplifying the function f .
- Such approximations are particularly powerful in presence of convex structures.
- Let us start with variational inference in the exponential family.

Exponential Families

- Suppose we have iid data x_1, \dots, x_n and we consider a collection of sufficient statistics $\{\phi_k(X)\}_k$

- The empirical expectations of these statistics are

$$\hat{\mu}_k = \frac{1}{n} \sum_i \phi_k(x_i)$$

- Q: Can we build a distribution $p(x)$ consistent with these empirical moments? i.e.

$$\mathbb{E}_{X \sim p(x)} \{\phi_k(X)\} = \hat{\mu}_k \text{ for all } k.$$

Exponential Families

- Suppose we have iid data x_1, \dots, x_n and we consider a collection of sufficient statistics $\{\phi_k(X)\}_k$

- The empirical expectations of these statistics are

$$\hat{\mu}_k = \frac{1}{n} \sum_i \phi_k(x_i)$$

- Q: Can we build a distribution $p(x)$ consistent with these empirical moments? i.e.

$$\mathbb{E}_{X \sim p(x)} \{\phi_k(X)\} = \hat{\mu}_k \text{ for all } k.$$

- In general, this is an underdetermined problem. How to choose wisely amongst all possible solutions?

Exponential Families and Maximum Entropy

- A reasonable choice is to consider the distribution with *maximum entropy* subject to the empirical moments:

$$p^* = \arg \max_p H(p) , \text{ s.t. } \mathbb{E}_p\{\phi_k(X)\} = \hat{\mu}_k \text{ for all } k.$$

Shannon Entropy: $H(p) = -\mathbb{E}\{\log(p)\}$.

Exponential Families and Maximum Entropy

- A reasonable choice is to consider the distribution with *maximum entropy* subject to the empirical moments:

$$p^* = \arg \max_p H(p) , \text{ s.t. } \mathbb{E}_p\{\phi_k(X)\} = \hat{\mu}_k \text{ for all } k.$$

Shannon Entropy: $H(p) = -\mathbb{E}\{\log(p)\}$.

- The general form of maximum entropy is

$$p(x) \propto \exp \left\{ \sum_k \lambda_k \phi_k(x) \right\}$$

λ_k : Lagrange multipliers adjusted such that $\mathbb{E}_p \phi_k(X) = \hat{\mu}_k$ for all k .

Exponential Families

- The exponential family associated with ϕ is defined as the parametric family:

$$p_\theta(x) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\} , \text{ with}$$

$$A(\theta) = \log \int \exp\{\langle \theta, \phi(x) \rangle\} dx \quad \text{log-partition function}$$

Exponential Families

- The exponential family associated with ϕ is defined as the parametric family

$$p_\theta(x) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\} , \text{ with}$$

$$A(\theta) = \log \int \exp\{\langle \theta, \phi(x) \rangle\} dx \quad \text{log-partition function}$$

- It is well defined for the family of parameters

$$\Omega = \{\theta ; A(\theta) < \infty\}$$

- Several well-known models belong to the exponential family:

- Energy based models
- Gaussian Mixtures
- Latent Dirichlet Allocation

Exponential Families

- **Proposition:** The log-partition function $A(\theta)$ satisfies

$$\frac{\partial A}{\partial \theta_k}(\theta) = \mathbb{E}_\theta\{\phi_k(X)\} = \int \phi_k(x)p_\theta(x)dx .$$

- $A(\theta)$ is convex in its domain Ω .

- Higher order derivatives always exist.

Legendre Transform

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a convex function.

The Legendre Transform f^* of f is defined as

$$f^*(u) = \sup_x (xu - f(x))$$

- f^* is the convex conjugate of f .

Legendre Transform

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a convex function.

The Legendre Transform f^* of f is defined as

$$f^*(u) = \sup_x (xu - f(x))$$

- f^* is the convex conjugate of f .
- Equivalent definition in the differentiable case:

f and g are Legendre transforms of each other if their first derivatives are inverses of each other:

$$\forall x, g'(f'(x)) = x , \quad \forall u, f'(g'(u)) = u .$$

- It follows that $f^{**} = f$

Conjugate Duality

- Conjugate duality representation of convex functions:

$$A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \}$$

canonical parameters \longleftrightarrow moment parameters

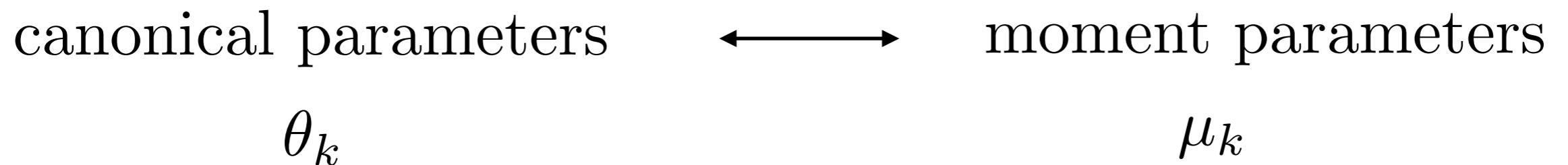
θ_k μ_k

- Q: How to interpret the dual conjugate?

Conjugate Duality

- Conjugate duality representation of convex functions:

$$A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \}$$



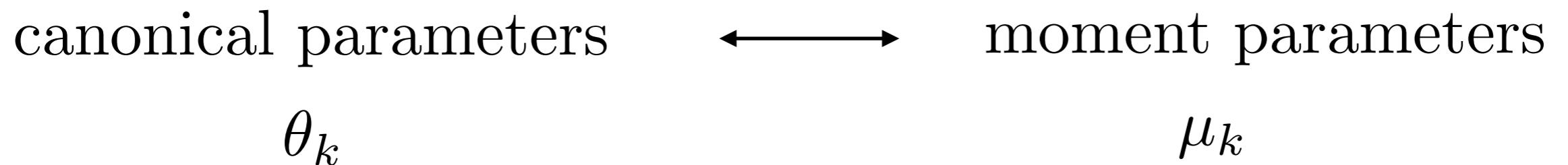
- Q: How to interpret the dual conjugate?

$A^*(\mu)$: Negative entropy of $p_{\theta(\mu)}$, where $p_{\theta(\mu)}$ is the exponential family distribution such that $\mathbb{E}_{\theta(\mu)} \phi(X) = \mu$.

Conjugate Duality

- Conjugate duality representation of convex functions:

$$A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \}$$



- Q: How to interpret the dual conjugate?

$A^*(\mu)$: Negative entropy of $p_{\theta(\mu)}$, where $p_{\theta(\mu)}$ is the exponential family distribution such that $\mathbb{E}_{\theta(\mu)} \phi(X) = \mu$.

- Variational representation:

$$A(\theta) = \sup_{\mu} \{ \langle \theta, \mu \rangle - A^*(\mu) \}$$

Variational Inference and Duality

- We derive the exact EM algorithm for exponential families with latent variables. Given observed variables \mathbf{Z} and latent variables \mathbf{X} , we consider

$$p_{\theta}(x, z) = \exp \{ \langle \theta, \phi(x, z) \rangle - A(\theta) \} , \text{ with}$$

$$A(\theta) = \log \int_{x,z} \exp \{ \langle \theta, \phi(x, z) \rangle \} dx dz$$

Variational Inference and Duality

- We derive the exact EM algorithm for exponential families with latent variables. Given observed variables \mathbf{Z} and latent variables \mathbf{X} , we consider

$$p_{\theta}(x, z) = \exp \{ \langle \theta, \phi(x, z) \rangle - A(\theta) \} , \text{ with}$$

$$A(\theta) = \log \int_{x,z} \exp \{ \langle \theta, \phi(x, z) \rangle \} dx dz$$

- Given observation $\mathbf{X} = x$, the posterior distribution is

$$p(z \mid x) = \frac{\exp \{ \langle \theta, \phi(x, z) \rangle \}}{\int \exp \{ \langle \theta, \phi(x, z') \rangle \} dz'} = \exp \{ \langle \theta \phi(x, z) \rangle - A_x(\theta) \}$$

$$A_x(\theta) = \log \int_z \exp \{ \langle \theta, \phi(x, z) \rangle \} dz$$

Variational Inference and Conjugate Duality

- The MLE for our parameters θ is obtained by maximizing the incomplete log-likelihood of the data:

$$\mathcal{L}(\theta, x) = \log \int_z \exp\{\langle \theta, \phi(x, z) \rangle - A(\theta)\} dz = A_x(\theta) - A(\theta) .$$

Variational Inference and Conjugate Duality

- The MLE for our parameters θ is obtained by maximizing the incomplete log-likelihood of the data:

$$\mathcal{L}(\theta, x) = \log \int_z \exp\{\langle \theta, \phi(x, z) \rangle - A(\theta)\} dz = A_x(\theta) - A(\theta).$$

- The variational representation gives

$$A_x(\theta) = \sup_{\mu_x} \{ \langle \theta, \mu_x \rangle - A_x^*(\mu_x) \}$$

$$A_x^*(\mu_x) = \sup_{\theta} \{ \langle \theta, \mu_x \rangle - A_x(\theta) \}$$

Variational Inference and Conjugate Duality

- The MLE for our parameters θ is obtained by maximizing the incomplete log-likelihood of the data:

$$\mathcal{L}(\theta, x) = \log \int_z \exp\{\langle \theta, \phi(x, z) \rangle - A(\theta)\} dz = A_x(\theta) - A(\theta).$$

- The variational representation gives

$$A_x(\theta) = \sup_{\mu_x} \{ \langle \theta, \mu_x \rangle - A_x^*(\mu_x) \}$$

$$A_x^*(\mu_x) = \sup_{\theta} \{ \langle \theta, \mu_x \rangle - A_x(\theta) \}$$

- It results in the lower-bound for the incomplete log-likelihood:

$$\mathcal{L}(\theta, x) \geq \langle \mu_x, \theta \rangle - A_x^*(\mu_x) - A(\theta) = \tilde{\mathcal{L}}(\mu_x, \theta)$$

Variational Inference and Conjugate Duality

- EM is thus a coordinate ascent on the lower bound:

$$\mu_x^{(t+1)} = \arg \max_{\mu_x} \tilde{\mathcal{L}}(\mu_x, \theta^{(t)}) \quad (\text{E step})$$

$$\theta^{(t+1)} = \arg \max_{\theta} \tilde{\mathcal{L}}(\mu_x^{(t+1)}, \theta) \quad (\text{M step})$$

Variational Inference and Conjugate Duality

- EM is thus a coordinate ascent on the lower bound:

$$\mu_x^{(t+1)} = \arg \max_{\mu_x} \tilde{\mathcal{L}}(\mu_x, \theta^{(t)}) \quad (\text{E step})$$

$$\theta^{(t+1)} = \arg \max_{\theta} \tilde{\mathcal{L}}(\mu_x^{(t+1)}, \theta) \quad (\text{M step})$$

- E step is called expectation because the maximizer of $\tilde{\mathcal{L}}(\mu_x, \theta)$ is, by duality, the expectation $\mu_x^{(t+1)} = \mathbb{E}_{\theta^{(t)}} \phi(x, Z)$

Variational Inference and Conjugate Duality

- EM is thus a coordinate ascent on the lower bound:

$$\mu_x^{(t+1)} = \arg \max_{\mu_x} \tilde{\mathcal{L}}(\mu_x, \theta^{(t)}) \quad (\text{E step})$$

$$\theta^{(t+1)} = \arg \max_{\theta} \tilde{\mathcal{L}}(\mu_x^{(t+1)}, \theta) \quad (\text{M step})$$

- E step is called expectation because the maximizer of $\tilde{\mathcal{L}}(\mu_x, \theta)$ is, by duality, the expectation $\mu_x^{(t+1)} = \mathbb{E}_{\theta^{(t)}} \phi(x, Z)$
- Also, because $\max_{\mu} \{\langle \mu_x, \theta^{(t)} \rangle - A_x^*(\mu_x)\} = A_x(\theta^{(t)})$, after each E step the inequality becomes an equality, thus M step increases log-likelihood.

Approximate Posterior Inference

- For most models, the posterior is analytically intractable:

$$p(z \mid x) = \frac{p(x \mid z)p(z)}{\int p(x \mid z')p(z')dz'}$$

- **Variational Bayesian Inference:** consider a parametric family of approximations $q(z \mid \beta)$ and optimize variational lower bound with respect to the variational parameters β .

Mean Field Variational Bayes

- Joint likelihood of observed and latent variables:
 $p(X, Z \mid \theta)$ θ : generative model parameters

- Let us consider a posterior approximation $q(z|\beta)$ of the form

$$q(z \mid \beta) = \prod_i q_i(z_i \mid \beta_i) \quad \beta: \text{Variational parameters}$$

- Mean-field approximation: we model hidden variables as being independent.

Mean Field Variational Bayes

- Joint likelihood of observed and latent variables:
 $p(X, Z | \theta)$ θ : generative model parameters

- Let us consider a posterior approximation $q(z|\beta)$ of the form

$$q(z | \beta) = \prod_i q_i(z_i | \beta_i) \quad \beta: \text{Variational parameters}$$

- Mean-field approximation: we model hidden variables as being independent.
- Corresponding lower-bound is given by

$$\log p(X | \theta) \geq \int q(z | \beta) \log \frac{p(x, z | \theta)}{q(z | \beta)} dz = \mathbb{E}_{q(z|\beta)} \{\log(p(X, Z | \theta))\} + H(q(z | \beta))$$

Mean Field Variational Bayes

- **Goal:** optimize lower-bound with respect to variational parameters.

Mean Field Variational Bayes

- **Goal:** optimize lower-bound with respect to variational parameters.
- As we have seen, this is equivalent to minimizing the divergence between true and approximate posterior:

$$\log p(X \mid \theta) = \tilde{\mathcal{L}}(\theta, \beta) + D_{KL}(q_\beta(z) \parallel p(z|x, \theta))$$

Mean Field Variational Bayes

- **Goal:** optimize lower-bound with respect to variational parameters.
- As we have seen, this is equivalent to minimizing the divergence between true and approximate posterior:

$$\log p(X \mid \theta) = \tilde{\mathcal{L}}(\theta, \beta) + D_{KL}(q_\beta(z) \parallel p(z|x, \theta))$$

- If $q(z \mid \beta)$ is a factorial distribution, the entropy term is tractable:

$$H(q(z|\beta)) = \sum_i H(q_i(z_i|\beta_i))$$

- Problematic term: $\nabla_\beta \mathbb{E}_{q(z|\beta)} \log p(X, Z|\theta)$

Mean Field Variational Bayes

- Denote

$$f(Z) = \log p(X, Z|\theta)$$

[Paiskey, Blei, Jordan, '12]

- Then

$$\begin{aligned}\nabla_{\beta} \mathbb{E}_{q(z|\beta)} f(Z) &= \nabla_{\beta} \int f(z) q(z|\beta) dz \\ &= \int f(z) \nabla_{\beta} q(z|\beta) dz \\ &= \int f(z) q(z|\beta) \nabla_{\beta} \log q(z|\beta) dz \\ &= \mathbb{E}_q \{ f(Z) \nabla_{\beta} \log q(z|\beta) \}\end{aligned}$$

Mean Field Variational Bayes

- Denote

$$f(Z) = \log p(X, Z|\theta)$$

[Paisley, Blei, Jordan, '12]

- Then

$$\begin{aligned}\nabla_{\beta} \mathbb{E}_{q(z|\beta)} f(Z) &= \nabla_{\beta} \int f(z) q(z|\beta) dz \\ &= \int f(z) \nabla_{\beta} q(z|\beta) dz \\ &= \int f(z) q(z|\beta) \nabla_{\beta} \log q(z|\beta) dz \\ &= \mathbb{E}_q \{ f(Z) \nabla_{\beta} \log q(z|\beta) \}\end{aligned}$$

- Stochastic approximation of

:

$$\nabla_{\beta} \mathbb{E}_{q(z|\beta)} f(Z) \approx \frac{1}{S} \sum_{s \leq S, z^{(s)} \sim q(z|\beta)} f(z^{(s)}) \nabla_{\beta} \log q(z^{(s)}|\beta)$$

Mean Field Variational Bayes

- The estimator of the gradient is unbiased, but it may suffer from large variance.
 - We may need a large number S of samples to stabilize the descent.
 - This estimator is also the basis of policy gradients in RL.
- Faster alternative?