Constrained Data-Fitters

Larry Samuelson, Jakub Steiner

Caltech, April 2024

homo economicus flawlessly

- forms Bayesian updates
- evaluates likelihood

machine learning: the two tasks can only be approximated

we: halfway between machine learning and economics

- we relax constraints enough

constrained-optimal models are often simple

Literature

variational Bayes methods: Jordan et al.'99, Kingma&Welling'13, Aridor, da Silveira&Woodford'24

- approximate Bayes' rule and maximum-likelihood estimation

misspecified learning: Berk'66, White'82, Esponda&Pouzo'16

- arises as a special case

causal networks: Pearl'09, Spiegler'16

- description of cognitive constraints

information design: Aumann&Maschler'95, Kamenica&Gentzkow'11, Caplin&Dean'13

- posterior approach

Generative Model

an agent holds a model $p(x, z) \in \Delta(X \times Z)$ of

- observable x
- latent z

economics:

- x is the signal (education level)
- z is the state (applicant's type)

machine learning:

- x is high-dimensional data input (job interview)
- z is a compressed representation of x (classification of the applicant)

Generative Model

an agent holds a generative model $p(x, z) \in \Delta(X \times Z)$ of

- observable x
- latent z

economics:

- x is the signal (education level)
- z is the state (applicant's type)

machine learning:

- x is high-dimensional data input (job interview)
- z is a compressed representation of x (classification of the applicant)

Recognition Model

the agent

- observes a realization x drawn from true process $q_0(x) \neq p(x)$
- forms update $q(z \mid x) \neq p(z \mid x)$

recognition model:

$$q(x, z) = q_0(x)q(z \mid x)$$

Constrained Updating

variational Bayes methods, Jordan et al.'99

updates solve

$$\begin{aligned} \max_{\left(\tilde{q}(z|x)\right)_x} \quad & \mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \mathrm{H}\left(\tilde{q}(x,z)\right) \\ \text{s.t.} \quad & \tilde{q}(x,z) \in \mathcal{Q} \end{aligned}$$

the maximizer: constrained updates

the value: constrained likelihood

Constrained Updating

variational Bayes methods, Jordan et al.'99

recognition model solves

$$\begin{aligned} \max_{\tilde{q}(x,z)} \quad & \mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \mathrm{H}\left(\tilde{q}(x,z)\right) \\ \text{s.t.} \quad & \tilde{q}(x,z) \in \mathcal{Q} \\ & \tilde{q}(x) = q_0(x) \end{aligned}$$

empirical constraint

Constrained Updating

variational Bayes methods, Jordan et al.'99

recognition model solves

$$\max_{\tilde{q}(x,z)} \quad \mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \mathrm{H}\left(\tilde{q}(x,z)\right)$$

$$\text{s.t.} \quad \tilde{q}(x,z) \in \mathcal{Q}$$

$$\tilde{q}(x) = q_0(x)$$

updating constraint

Constrained Updating

variational Bayes methods, Jordan et al.'99

recognition model solves

$$\begin{aligned} \max_{\tilde{q}(x,z)} \quad & \mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \mathrm{H}\left(\tilde{q}(x,z)\right) \\ \text{s.t.} \quad & \tilde{q}(x,z) \in \mathcal{Q} \\ & \tilde{q}(x) = q_0(x) \end{aligned}$$

reconstruction term

Constrained Updating

variational Bayes methods, Jordan et al.'99

recognition model solves

$$\begin{aligned} \max_{\tilde{q}(x,z)} \quad & \mathrm{E}_{\tilde{q}(x,z)} \ln p(\hat{x}, \hat{z}) + \mathrm{H}\left(\tilde{q}(x,z)\right) \\ \text{s.t.} \quad & \tilde{q}(x,z) \in \mathcal{Q} \\ & \tilde{q}(x) = q_0(x) \end{aligned}$$

regularization term

Constrained Updating

variational Bayes methods, Jordan et al.'99

recognition model solves

$$\begin{aligned} \min_{\tilde{q}(x,z)} \quad & \mathrm{KL}\left(\tilde{q}(x,z) \,\|\, p(x,z)\right) \\ \text{s.t.} \quad & \tilde{q}(x,z) \in \mathcal{Q} \\ & \tilde{q}(x) = q_0(x) \end{aligned}$$

Some Updating Constraints

no constraint: $\mathcal{Q} = \Delta(X \times Z)$

- Bayesian updates, $q(z \mid x) = p(z \mid x)$
- unconstrained likelihood, $\mathrm{E}_{q_0(x)} \ln p(\hat{x}) + \text{const.}$

analogy-based constraint: $q(z \mid x)$ measurable w.r.to a partition of X

causal constraint; e.g.:

- $z = (z_1, z_2)$
- q must comply with directed acyclical graph $z_1 \leftarrow x \rightarrow z_2$
- $\Leftrightarrow$ factorization constraint $q(x, z_1, z_2) = q(x)q(z_1 \mid x)q(z_2 \mid x)$

Likelihood Evaluation

sample $(x_1, \ldots, x_n)$

via marginalization:

- $p(x) = \sum_z p(x, z)$
- $\ell = \prod_i p(x_i)$

via sample extension:

- extended sample $(x_i, z_i)_{i=1}^n$
- frequencies of (x, z): frequencies of x observed & $z \mid x \sim p(z \mid x)$
- $\ell = \prod_i p(x_i, z_i) \times$ no. of distinct permutations

updating and fit evaluation are related

Constrained updating

estimate frequencies $q(x, z)$ in the extended sample

$$\begin{aligned} \max_{\tilde{q}(x,z)} \quad & p\text{-likelihood} \\ \text{s.t.} \quad & \tilde{q}(x) = q_0(x) \\ & \tilde{q}(x,z) \in \mathcal{Q} \end{aligned}$$

Estimation

p-likelihood of an extended sample with frequencies $q(x, z)$ is

$$\prod_{i=1}^{n} p(x_i, z_i) = \prod_{x,z} p(x, z)^{q(x,z)n}$$

p-likelihood of all such extended samples

$$\ell_n(q) := \prod_{x,z} p(x, z)^{q(x,z)n} \times \mathcal{N}_n(q)$$

the estimate:

$$q_n(x, z) \in \arg\max_{\tilde{q} \in \mathcal{Q}_n} \ell_n(q)$$

$x \in \{r, b\}$

$z \in \{0, 1\}$

$x^4 = rbrb$

consider $q(x, z)$ uniform on $\{r, b\} \times \{0, 1\}$

four possible extended samples:

x^4	r	b	r	b
z^4	0	0	1	1
z^4	1	0	0	1
z^4	0	1	1	0
z^4	1	1	0	0

Limit

let $\mathcal{Q}_n$ approximate $\mathcal{Q}$ details

proposition

$$
\begin{aligned}
q_n(x,z) &\quad\rightarrow\quad \text{recognition model } q(x,z) \\
\frac{1}{n}\ln \ell_n(q_n) &\quad\rightarrow\quad \text{constrained likelihood} + \text{const.}
\end{aligned}
$$

because

$$
\begin{aligned}
\frac{1}{n}\ln & \qquad \prod_{x,z} p(x,z)^{q(x,z)n} \quad \times \quad \mathcal{N}_n(q) \\
\rightarrow & \qquad \mathrm{E}_{q(x,z)} \ln p(\hat{x},\hat{z}) \quad + \quad \mathrm{H}\left(q(x,z)\right) - \mathrm{H}\left(q_0(x)\right)
\end{aligned}
$$

Approximate Maximum Likelihood Estimation

variational autoencoder, Kingma&Welling'13

principle

Choose the generative model that maximizes constrained likelihood.

$$
\begin{aligned}
\min_{\tilde{p}(x,z),\tilde{q}(x,z)} \quad & \mathrm{KL}\left(\tilde{q}(x,z) \parallel \tilde{p}(x,z)\right) \\
\text{s.t.} \quad & \tilde{p}(x,z) \in \mathcal{P} \\
& \tilde{q}(x,z) \in \mathcal{Q} \\
& \tilde{q}(x) = q_0(x)
\end{aligned}
$$

Causal Constraint
example

$z = (z_1, z_2)$

recognition model restricted to a chain: $x \to z_1 \to z_2$

$\mathcal{P}$ has unconstrained margin:

- all $p(z)$ are feasible
- a constraint on $\big(p(x \mid z)\big)_z$ independent of $p(z)$

deterministic collapse

The agent forms a partially deterministic model:

$$z_2 = d(z_1)$$

a.s. under both p and q, for some deterministic function d.

Rational Expectations
definition

we identify rational expectations with

$$p(z) = \mathrm{E}_{q_0(x)}\, q(z \mid \hat{x}) \equiv q(z)$$

in general, our agent will not form RE because she

- isn't Bayes' rational
- is misspecified

yet, under a condition, the agent forms RE

Rational Expectations

result

proposition

If $\mathcal{P}$ has unconstrained margin, then agent forms rational expectations.

proof: optimize over $\tilde{p}(z)$,

$$\mathrm{KL}\left(q(x,z) \parallel \tilde{p}(x,z)\right) = \mathrm{KL}\left(q(z) \parallel \tilde{p}(z)\right) + \sum_{z} q(z)\,\mathrm{KL}\left(q(x \mid z) \parallel p(x \mid z)\right)$$

Discussion

standard Bayes' plausibility is forced by the Bayes' law

- it can fail in our framework, but holds at the optimum

a popular non-Bayesian intuition in support of RE:

- systematically surprised agent should adjust her prior
- indeed, $p(z)$ is chosen to match $q(z)$

Posterior Approach
information design

posterior representation: $q(z)$, $\big(q(x \mid z)\big)_z$, and $\big(p(x \mid z)\big)_z$

- specifies both models $p(x,z)$ and $q(x,z)$

lemma: posterior-separable objective

If $\mathcal{P}$ has unconstrained margin, then the model-fitting problem becomes

$$\begin{aligned}
\max_{\tilde{q}(z),(\tilde{q}(x|z))_z,(\tilde{p}(x|z))_z} \quad & \mathrm{E}_{\tilde{q}(z)}\left[\mathrm{E}_{\tilde{q}(x|\hat{z})} \ln \tilde{p}(\hat{x} \mid \hat{z}) + \mathrm{H}\left(\tilde{q}(x \mid \hat{z})\right)\right] \\
\text{s.t.} \quad & (\tilde{p}(x \mid z))_z \in \mathcal{P}' \\
& \tilde{q}(z)\tilde{q}(x \mid z) \equiv \tilde{q}(x,z) \in \mathcal{Q} \\
& \mathrm{E}_{\tilde{q}(z)}\, \tilde{q}(x \mid z) = q_0(x).
\end{aligned}$$

Posterior Approach
information design

posterior representation: $q(z)$, $\big(q(x \mid z)\big)_z$, and $\big(p(x \mid z)\big)_z$

- specifies both models $p(x, z)$ and $q(x, z)$

lemma: posterior-separable objective

If $\mathcal{P}$ has unconstrained margin, then the model-fitting problem becomes

$$\begin{aligned} \max_{\tilde{q}(z), (\tilde{q}(x|z))_z, (\tilde{p}(x|z))_z} \quad & \mathrm{E}_{\tilde{q}(z)} \mathrm{KL}\left(\tilde{q}(x \mid \hat{z}) \parallel \tilde{p}(x \mid \hat{z})\right) \\ \text{s.t.} \quad & (\tilde{p}(x \mid z))_z \in \mathcal{P}' \\ & \tilde{q}(z)\tilde{q}(x \mid z) \equiv \tilde{q}(x, z) \in \mathcal{Q} \\ & \mathrm{E}_{\tilde{q}(z)} \tilde{q}(x \mid z) = q_0(x). \end{aligned}$$

Deterministic Collapse
proof

recall the chain constraint $\mathcal{Q}$: $x \to z_1 \to z_2$

- equivalent to $q(x \mid z_1, z_2) = q(x \mid z_1)$

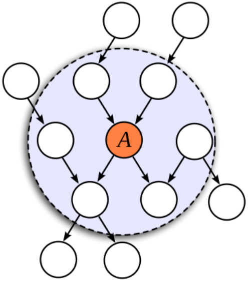
for each z_1, optimize over $q(z_2 \mid z_1)$

z_2 affects $\mathrm{E}_{\tilde{q}(x|z_1)} \ln p(\hat{x} \mid z_1, z_2) + \mathrm{H}\left(\tilde{q}(x \mid z_1)\right)$ only via p

deterministically pick $z_2^*(z_1)$ that maximizes this

Markov Boundary

Pearl '88



Markov boundary of A: minimal set that contains all information about A

e.g. in $x \rightarrow z_1 \rightarrow z_2$

- z_1 is in the Markov boundary of x
- z_2 isn't

General Simplicity Result

fix DAG

z^B – the latent variables from Markov boundary of x

- $q(x \mid z)$ depends only on z^B

say q' is simpler than q if

- $q'(x, z^B) = q(x, z^B)$, and
- $z^{-B} \mid z^B$ is deterministic under q'

$\mathcal{Q}$: q compatible with the DAG and any q' simpler than q

deterministic collapse

A solution exists such that latent variables from outside of the Markov boundary of x are deterministic functions of the variables from within the boundary.

Two Frictions

		$\mathcal{P}$	
		Well-specified	**Miss-specified**
$\mathcal{Q}$	**Bayes' Rationality**	Wald'49	Berk'66
	Updating Friction	model-fitting	model-fitting

Information vs Moment Projection

moment projection: sample $\rightarrow$ model

White'82/Berk'66:

agent observes sample and chooses model $\tilde{p}(y) \in \mathcal{P}$

$$\min_{\tilde{p} \in \mathcal{P}} \mathrm{KL}\left(q_0(y) \parallel \tilde{p}(y)\right)$$

information projection: model $\rightarrow$ sample

Sanov's Theorem:

agent holds model $p(y)$ and reasons about sample

$$\min_{\tilde{q} \in \mathcal{Q}} \mathrm{KL}\left(\tilde{q}(y) \parallel p(y)\right)$$

Example: Analogy-Based Reasoning

a measurability constraint on conditional distributions

moment projection $\Rightarrow$ arithmetic mean (Jehiel'05)

$$p(z \mid x) \propto \sum_{\tilde{x} \in X(x)} q_0(\tilde{x}) q_0(z \mid \tilde{x})$$

information projection $\Rightarrow$ geometric mean

$$q(z \mid x) \propto \left(\prod_{\tilde{x} \in X(x)} p(z \mid \tilde{x})^{q_0(\tilde{x})} \right)^{\frac{1}{q_0\left(X_{k(x)}\right)}}$$

White/Berk As a Special Case

what model $p(x)$ of the observable variable the agent chooses?

proposition

If updating is unconstrained, then $p(x)$ is the moment projection

$$p(x) \in \underset{\tilde{p}(x) \in \mathcal{P}'}{\arg\min} \, \mathrm{KL}\left(q_0(x) \parallel \tilde{p}(x)\right).$$

follows from the chain rule

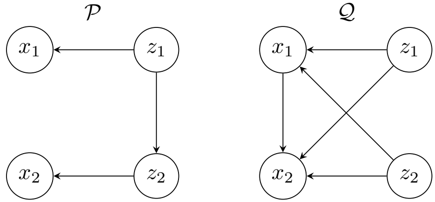
$$\mathrm{KL}\left(\tilde{q}(x,z) \parallel \tilde{p}(x,z)\right) = \mathrm{KL}\left(q_0(x) \parallel \tilde{p}(x)\right) + \sum_x q_0(x) \, \mathrm{KL}\left(\tilde{q}(z \mid x) \parallel \tilde{p}(z \mid x)\right)$$

Simple Model Preferred for Constrained Updating

example

$x = (x_1, x_2)$ and $z = (z_1, z_2)$

the true process $q_0(x_1, x_2)$ exhibits correlation



agent is well-specified

- $\Rightarrow$ learns the true process q_0 if updates are unconstrained

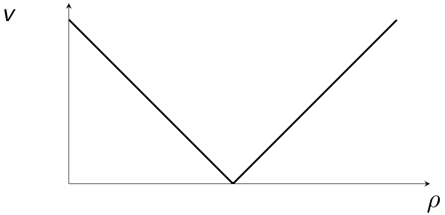
the updating constraint

- $\Rightarrow$ optimal correlation neglect $p(x_1, x_2) = p(x_1)p(x_2)$ proof

Rational Inattention

- payoff state $x \sim q_0(x)$
- agent chooses experiment $q(z \mid x)$
- maps the observed signal to action a
- maximizes $\mathrm{E}\, u(a, x) + \mathrm{E}\,\mathrm{H}\left(q(x \mid z)\right)$

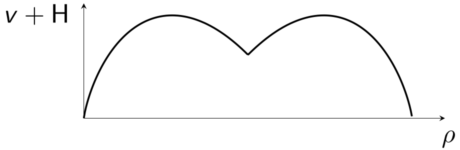
plot $\rho \mapsto \max_a \mathrm{E}_{\rho(x)}\, u(a, \hat{x})$



Rational Inattention

- payoff state $x \sim q_0(x)$
- agent chooses experiment $q(z \mid x)$
- maps the observed signal to action a
- maximizes $\mathrm{E}\, u(a, x) + \mathrm{E}\,\mathrm{H}\left(q(x \mid z)\right)$

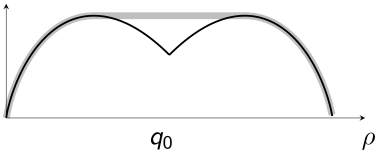
plot $\rho \mapsto \max_a \mathrm{E}_{\rho(x)}\, u(a, \hat{x}) + \mathrm{H}\left(\rho\right)$



Rational Inattention

- payoff state $x \sim q_0(x)$
- agent chooses experiment $q(z \mid x)$
- maps the observed signal to action a
- maximizes $\mathrm{E}\, u(a, x) + \mathrm{E}\, \mathrm{H}\left(q(x \mid z)\right)$

find the optimal posteriors



Connection

$\mathcal{P}$ and $\mathcal{Q}$ are posterior separable if

- p and q are feasible iff $p(x \mid z) \in \bar{\mathcal{P}}$ and $q(x \mid z) \in \bar{\mathcal{Q}}$

primitive distributions are "actions" $\bar{\mathcal{P}} = \{p_a(x)\}_a$

writing $\ln p_a(x) = u(a, x)$, our problem becomes the RI problem:

$$\max \qquad \mathrm{E}\left[u\left(a, z\right) + \mathrm{H}\left(\tilde{q}(x \mid \hat{z})\right)\right]$$

with additional constraint: posteriors $\in \bar{\mathcal{Q}}$

Concavification of the Augmented Value Function

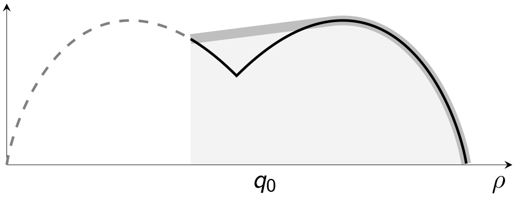
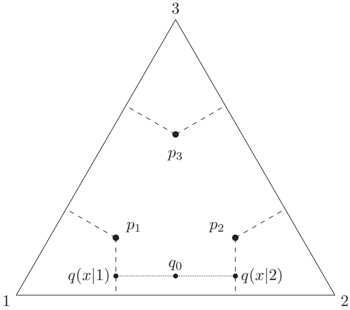


Illustration
Matysková&Montes'23

no updating constraint



generative model employs 1, 2, or 3 primitive distributions

- accompanied by a recognition model of the same complexity

Base-Rate Neglect

comparative statics w.r.to true process

local invariance

Let true process $q_0^*(x)$ induce posteriors by $p^*(x \mid z)$ and $q^*(x \mid z)$.
For all processes $q_0(x)$ in the convex hull of $\left(q^*(x \mid z)\right)_z$:

$$\begin{aligned} p(x \mid z) &= p^*(x \mid z) \\ q(x \mid z) &= q^*(x \mid z). \end{aligned}$$

Hallucination

optimal recognition model may hallucinate:

- there may exist z and z' such that

$$\begin{aligned} p(x \mid z) &= p(x \mid z') \\ q(x \mid z) &\neq q(x \mid z') \end{aligned}$$

this cannot happen when $\bar{\mathcal{Q}}$ is convex

- akin to the recommendation lemma in RI
- beneficial randomization over $q(x \mid z)$

Conclusion

machine learning $\rightarrow$ economics:

- updating and likelihood evaluation are hard
- two distinct statistical models are handy
- tractable constrained updating and model-fitting problems

economics $\rightarrow$ machine learning:

- relaxed constraints may generate solutions with interesting structure
- optimal models are often simple

Approximation

correspondence $\mathcal{Q}(\theta)$, $\theta \in [0,1]$

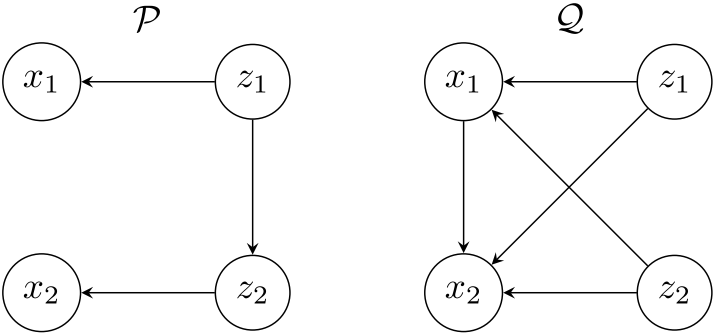
$\mathcal{Q}(0) = \mathcal{Q} \cap \left\{ \tilde{q}(x,z) : \tilde{q}(x) = q_0(x) \right\}$

$\mathcal{Q}(\theta) = \mathcal{Q}^{\lfloor \frac{1}{\theta} \rfloor}$ for $\theta > 0$

continuity at $\theta = 0$

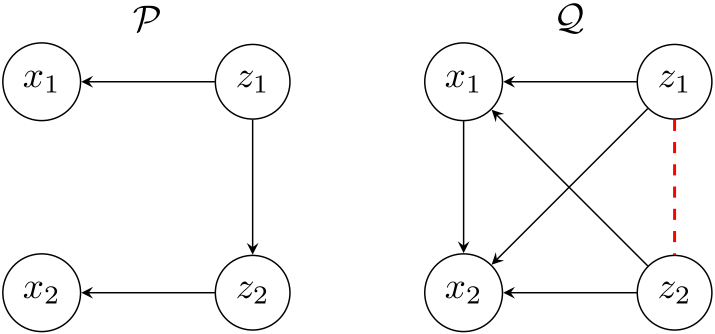
back

Proof: Correlation Neglect



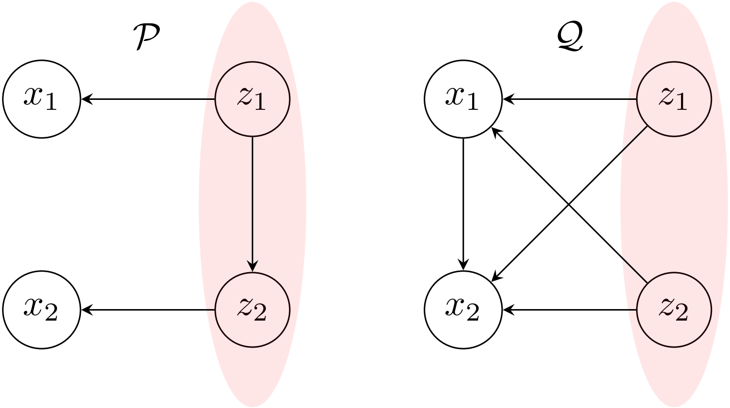
back

Proof: Correlation Neglect



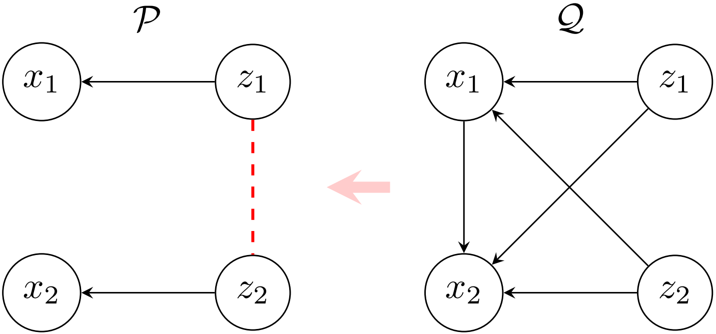
back

Proof: Correlation Neglect



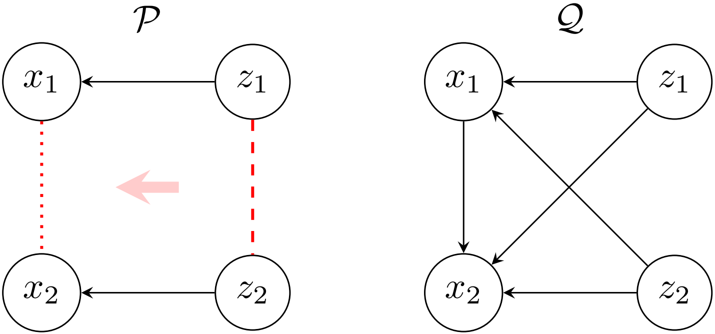
back

Proof: Correlation Neglect



back

Proof: Correlation Neglect



back