Progress in Normalizing Flows for 4d Gauge Theories

Ryan Abbott

MIT

July 25, 2024

Collaborators

Universität Bern — AEC Albert Einstein Center for Fundamental Physics

Massachusetts Institute of Technology


P. Shanahan

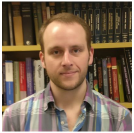

D. Boyda


R. Abbott


J. Urban


F. Romero-López

Fermilab

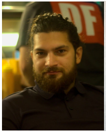

D. Hackett

DeepMind


S. Racanière


D. Rezende


A. Razavi


A. Botev


A. Matthews

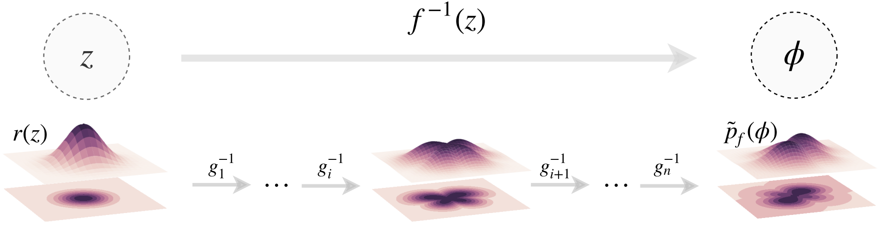
Universität Bern — AEC Albert Einstein Center for Fundamental Physics → The University of Edinburgh


G. Kanwar

New York University → Harvard


M. Albergo

Normalizing flows



[Albergo et al., 1904.12072]

- Learned change of variables f maps density $r(z)$

$$q(\phi) = |\det J_f(f(\phi))| r(f(\phi))$$

- $r(z), f^{-1}(z), |\det J_f(z)|$ tractable $\implies$ $q(\phi)$ tractable
- Given (known) target $p(\phi)$, train f so $q \approx p$
  - Can apply corrections for exact/unbiased sampling

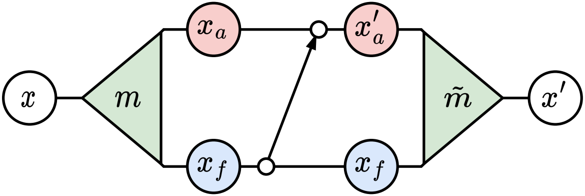
Normalizing flows & QCD

- Modern effort began w/ scalar fields [Albergo et al., 1904.12072]
- Required significant effort to get to QCD
  - Working with $\mathrm{U}(1)$ & SU(3), gauge symmetry, pseudofermions, ...
- Have tools for QCD [Abbott et al., 2208.03832]
- Outline today
  - More recent work on improving models
  - Scaling & Aurora (supercomputer)
  - Novel applications past accelerated sampling (Fernando)

Model improvements

Model improvements

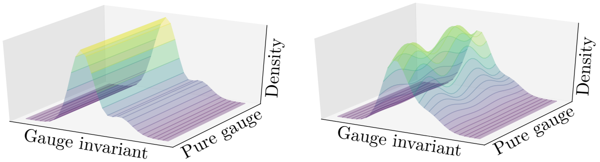
- Two main architectures: spectral & residual
  - See [Abbott et al, 2305.02402]
  - Both based on active/frozen split
- Many improvements to both
  - Diagonal features, learned active loops, initialization, ...
  - General theme: more gauge equivariant information
    - E.g. convolutions → parallel transport



Gauge Symmetry and Sampling

Gauge transformation

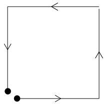
- Gauge symmetry $\Longrightarrow$ $p(\Omega \cdot U) = p(U)$
- Model gauge invariance: $q(\Omega \cdot U) = q(U)$
- Achieve with 2 conditions:
  - Prior gauge invariance: $r(\Omega \cdot U) = r(U)$
  - Gauge Equivariance: $f(\Omega \cdot U) = \Omega \cdot f(U)$



Spectral Flows

[Boyda et al., 2008.05456]

- Transform "active loop" (e.g. untraced plaquette $P_{\mu\nu}$)
- Under gauge transformation $\Omega(x) \in \mathrm{SU}(N)$



$$(\Omega \cdot P)_{\mu\nu}(x) = \Omega(x) P_{\mu\nu}(x) \Omega(x)^\dagger$$

- Given $h : \mathrm{SU}(N) \to \mathrm{SU}(N)$, transform U_μ so $P_{\mu\nu} \mapsto h(P_{\mu\nu})$

$$f(U_\mu) = h(P_{\mu\nu}) P_{\mu\nu}^\dagger U_\mu$$

- Gauge equivariance $\iff$ *conjugation equivariance*:

$$h(\Omega P \Omega^\dagger) = \Omega h(P) \Omega^\dagger$$

Spectral Flows

Goal: $h(\Omega X \Omega^\dagger) = \Omega h(X) \Omega^\dagger$

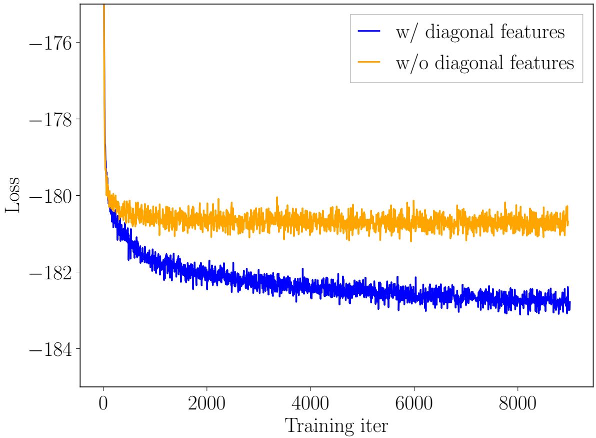
- Used for transforming active loop (plaquette, 2×1 loop, etc.)
- Conjugation invariant data $\Leftrightarrow$ eigenvalues
- Diagonalize $P \in \mathrm{SU}(N)$ via eigenbasis V:

$$P = V \begin{pmatrix} e^{i\theta_1} & & \\ & \ddots & \\ & & e^{i\theta_N} \end{pmatrix} V^\dagger \mapsto V \begin{pmatrix} e^{i\theta'_1} & & \\ & \ddots & \\ & & e^{i\theta'_N} \end{pmatrix} V^\dagger$$

- Define $h : \mathrm{SU}(N) \to \mathrm{SU}(N)$ by action on $\{\theta_1, \ldots, \theta_N\}$
  - Need to be careful about order $\Rightarrow$ choose canonical order
  - Note: θ_k not independent, $\prod_k e^{i\theta_k} = \det X = 1 \Rightarrow$ remove θ_N

Diagonal Features

- Eigenvectors V contain gauge-invariant information
  - E.g. $\mathrm{diag}(V^\dagger W V)$, $W =$ (frozen) Wilson loop
  - Use same canonical order as for eigenvalues

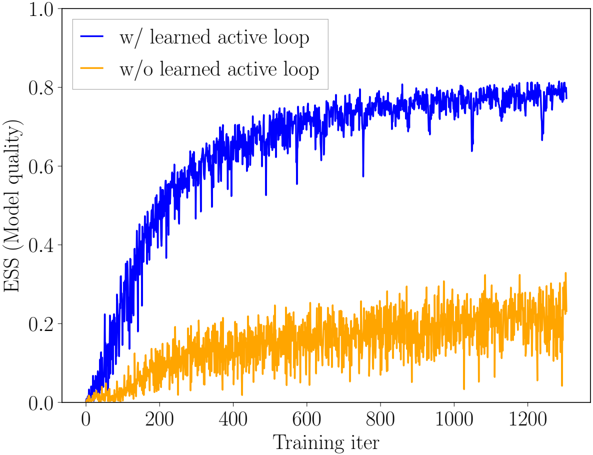


Learned active loops

- Usually use *fixed* active loop in each layer
  - E.g. plaquette, 2×1 loop
- Idea: use learned linear combination of possible loops
  - Constructed out of parallel transport + sitewise linear combinations
  - Similar to gauge-equivariant networks [Favoni et al., 2012.12901]
  - Project to SU(N) w/ polar projection $\mathcal{P}(M) = M(M^\dagger M)^{-1/2}$

Results

- Small test on 4^4 lattice, $\beta = 2$, 4d SU(3)



Scaling & Aurora

Comments on Scaling

- Reference: [Abbott et al., 2211.07541]
- Scaling depends strongly every aspect of the model
  - E.g. use of flow, architecture choices, training choices
  - Makes extrapolating beyond any particular choice difficult

Use of Flow

- Direct Sampling (Independence Metropolis)
- HMC on trivialized distribution [Lüscher 0907.5491]
- Generalize proposal distribution [Foreman et al., 2112.01582]
- Subdomain updates [Finkenrath, 2201.02216]
- Stochastic Normalizing Flows [Wu et al. 2002.0670]
- CRAFT [Matthews et al. 2201.13117]

Comments on Scaling

- Reference: [Abbott et al., 2211.07541]
- Scaling depends strongly every aspect of the model
  - E.g. use of flow, architecture choices, training choices
  - Makes extrapolating beyond any particular choice difficult

Architecture Choices

- Choice of coupling layers (spectral, residual, continuous)
- Choice of Neural networks (CNN, fully-connected, gauge-equivariant)
  - Gauge-equivariant networks [Favoni et al., 2012.12901]
- Choice of invariant context passed to networks
- Size of model (# layers, NN sizes)

Comments on Scaling

- Reference: [Abbott et al., 2211.07541]
- Scaling depends strongly every aspect of the model
  - E.g. use of flow, architecture choices, training choices
  - Makes extrapolating beyond any particular choice difficult

Training Choices

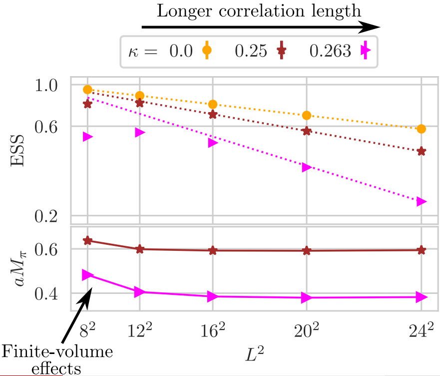
- Optimizer (Adam, SGD, higher-order optimizers)
- Choice of Loss (reverse/forward KL, MSE, ...)
- Computation of gradients (path gradients/control variates)
- Hyperparameter choices (batch size, learning rate)
  - Hyperparameter scheduling
- Volume chosen for training

Exponential Volume Scaling

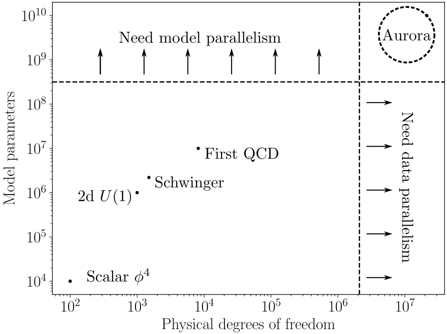
- For $L/\xi \gg 1$, $\xi =$ correlation length, volume transfer

$$ESS(V) = ESS(V_0)^{V/V_0}$$

- Prevents *direct sampling* in thermodynamic limit $L/\xi \to \infty$
  - Does not apply to continuum limit $L/\xi \sim m_\pi L$ fixed, $\xi/a \to \infty$
  - Typically $4 \lesssim m_\pi L \lesssim 10 \implies$ no in principle issue



Scaling On Aurora



- Aurora is an exascale machine at Argonne
- Significant software effort
  - Porting/checking code on Intel GPUs ✓
  - Distributing model + fields over multiple GPUs ✓
    - Note: training is very memory intensive
  - Model scaling to $O(10,000)$ GPUs – ongoing

Scaling on Aurora (continued)

- Significantly larger models, $\sim 10^9$–10^{10} parameters
  - Current models $\sim 10^6$–10^7 parameters
- Target: dynamical QCD, moderate size lattices
- Note: scaling ML models is highly nonintuitive, context-dependent
  - See [Abbott et al., 2211.07541] for a full discussion

GPT-1 (117 million parameters)
Lattice QCD is on and in the bag's not mine, "ben said. he was lying on the couch, . . .

GPT 3.5 ($\sim$ 175 billion parameters)
Lattice QCD is a numerical approach used in theoretical physics to study the strong interaction between quarks and gluons, which are the fundamental constituents of protons, neutrons, and other hadrons.

Conclusions

- Many improvements for 4d SU(3) flows
  - E.g. diagonal features, learned active loops
- Upcoming/ongoing scaling on Aurora

Conclusions

- Many improvements for 4d SU(3) flows
    - E.g. diagonal features, learned active loops
- Upcoming/ongoing scaling on Aurora
- Thanks! Questions?

Massachusetts Institute of Technology

AURORA | EARLY SCIENCE PROGRAM FOR DATA & LEARNIN

Backup

Unbiased sampling

- Independence Metropolis: accept $\phi \to \phi' \sim q(\phi')$ with probability

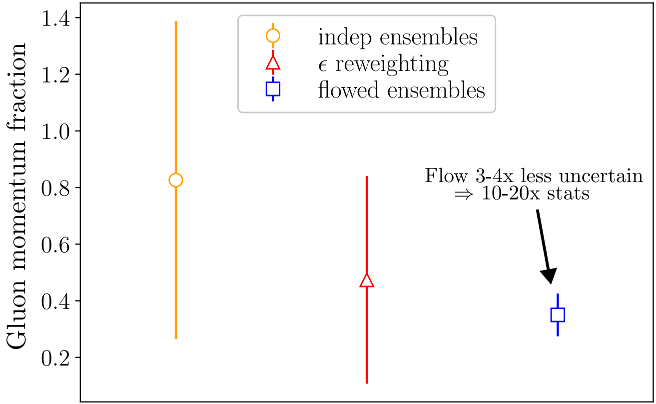
$$P_{\text{accept}}(\phi \to \phi') = \min\left(1, \frac{p(\phi')}{p(\phi)} \frac{q(\phi)}{q(\phi')}\right)$$

- Hybrid methods
  - Alternate HMC/flow updates
  - HMC on trivialized distribution [Lüscher 0907.5491]
  - Subdomain updates [Finkenrath, 2201.02216]
  - CRAFT/Annealed Importance Sampling [Matthews et al. 2201.13117]
  - ...

Novel applications of flows

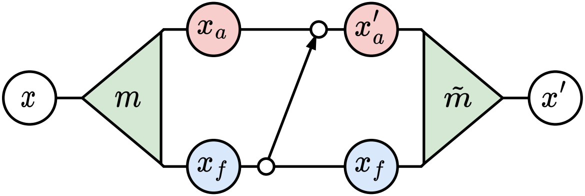
[Abbott et al., 2401.10874]

- If $f \approx$ identity (can force), then $f(U)$ and U are correlated
  - $\implies$ correlated differences, improved uncertainties
  - E.g. Feyman-Hellman, continuum limit
- Feynman-Hellman example:



SU(N)-Equivariant Flows

- Two types here
  - Spectral flows - transform untraced plaquettes
    - Reference: [Boyda et al., 2008.05456]
  - Residual flows - parametrized Wilson flow/stout smearing step
    - Reference: [Abbott et al., 2304.XXXXX] (to appear)
- Both based on active/frozen split



Residual Flows

- Inspired by Lüscher's trivializing map [Lüscher 0907.5491]
- Transform active links via

Lie-algebra-valued derivative

$$U_\mu(x) \mapsto e^{i\epsilon\partial_{x,\mu}\phi(U)} U_\mu(x)$$

- Gauge-invariant "potential" $\phi(U)$
  - Example: $\phi(U) \propto S_{\text{Wilson}}(U) \implies$ Wilson flow/stout smearing
  - More complex:

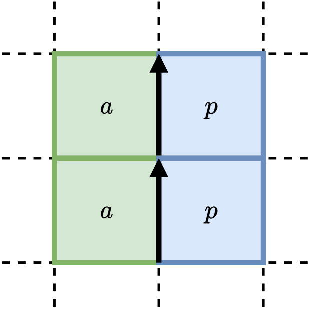
$$\phi(U) = \sum_x \sum_{\mu\neq\nu} c_{\mu\nu}(x; U_{\text{frozen}}) \mathrm{Re}\,\mathrm{Tr}(P_{\mu\nu})$$

- Small but finite ϵ for invertibility ($\epsilon \lesssim 1/8$)

Spectral vs Residual Flows

Spectral flows

- Transform plaquettes
- Limited by passive plaquettes



Residual flows

- Update links
- Denser active mask
- Limited by step size
- Harder to invert
    - Require fixed-point iteration

Continuous Flows

[Bacchio et al. 2212.08469]

- Continuous time
- Unmasked
- Requires ODE integration

Fermions

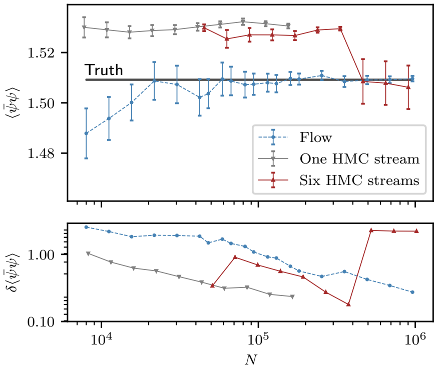
Fermion target:

$$p(U) \propto e^{-S_G[U]} \det M[U]$$

Methods:

- Compute $\det M$ directly
  - Simple, but not scalable
- Estimate $\det M$
  - E.g. pseudofermions

Schwinger model at criticality



[Albergo et al. 2202.11712]

Autoregressive Pseudofermion modeling

Target Distributions:

- Marginal:

$$p_m(U) = e^{-S_G(U)} \det M[U]$$

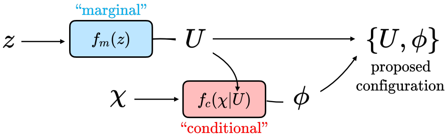
- Conditional:

$$p_c(\phi \mid U) \propto \frac{1}{\det M[U]} e^{-\phi^\dagger M^{-1} \phi}$$

- Joint:

$$p_{\text{joint}}(U, \phi) = p_c(\phi \mid U) p_m(U) = e^{-S_G(U) - \phi^\dagger M^{-1} \phi}$$

Models:



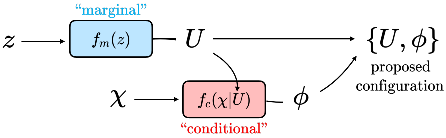
Prior:

- Gauge $z \sim$ Haar, heatbath, ...
- Pseudofermion $\chi \sim e^{-\chi^\dagger \chi}$

[Albergo et al., 2106.05934]
[Abbott et al., 2207.0945]

Conditional Model (2 Flavor Theory)

[Albergo et al., 2106.05934]
[Abbott et al., arxiv:2207.0945]



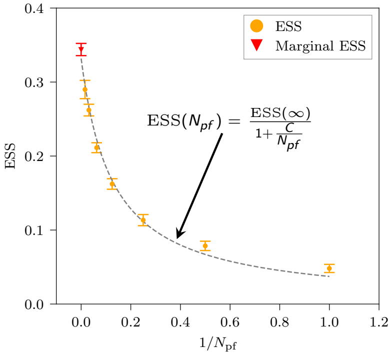
- Prior $\chi \sim e^{-\chi^\dagger \chi}$
- Target $\phi \sim \frac{1}{\det(DD^\dagger)} e^{-\phi^\dagger (DD^\dagger)^{-1} \phi}$
- *Optimal* model: $\phi = f_c(\chi \mid U) = D[U]\chi$
  - But $\det J = \det DD^\dagger$ not tractable
- Estimate optimal model with tractable (gauge-equivariant) layers

$$\phi_a(x) \mapsto A[U](x)\phi_a(x) + B[U](x,y)\phi_f(y)$$
$$\phi_f(x) \mapsto \phi_f(x)$$

- $A[U], B[U]$: (learned) linear operators

Improving Pseudofermion Models

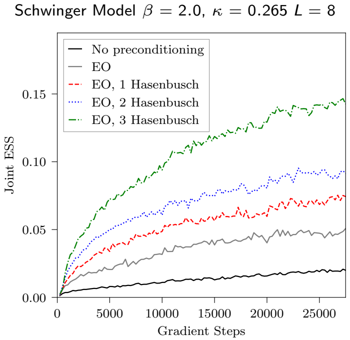
- More pseudofermion draws
  - Improve for fixed model



(ESS $\sim$ acceptance rate)

- Even/Odd preconditioning
- Hasenbusch factorization

$$\det(M) = \frac{\det(M)}{\det(M+\mu)} \det(M+\mu)$$



Training Marginal Models

- Stochastic derivative estimate:

$$\begin{aligned} \nabla \log \det M &= \operatorname{Tr} \nabla \log M \\ &= \operatorname{Tr}\left[M^{-1} \nabla M\right] \\ &= \mathbb{E}_{\chi \sim e^{-\chi^\dagger \chi}}\left[\chi^\dagger M^{-1} \nabla M \chi\right] \end{aligned}$$

- Requires 1 inversion/sample $\chi^\dagger M^{-1}$
- Does not give access to density

Unbiased sampling

- Define reweighting factors $w(\phi)$:

$$w(\phi) = \frac{p(\phi)}{q(\phi)}$$

- Reweighting/Importance Sampling:

$$\int \mathrm{d}\phi \, p(\phi)\mathcal{O}(\phi) = \int \mathrm{d}\phi \, w(\phi)\tilde{p}_f(\phi)\mathcal{O}(\phi) \approx \frac{1}{N} \sum_{\phi \sim q(\phi)} w(\phi)\mathcal{O}(\phi)$$

- Metropolis Hastings: accept $\phi \to \phi'$ with probability

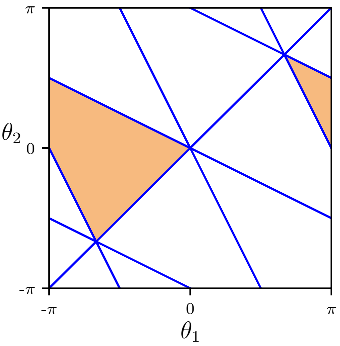
$$\min\left(1, \frac{w(\phi')}{w(\phi)}\right)$$

Spectral Flows (continued)

Goal: Permutation equivariant flow

- Perform *maximal torus flow* on $\{\theta_i\}$
- Choose (arbitrary) canonical cell
- Use order of eigenvalues
  - Canonicalization $\sim$ sorting
- In canonical cell: use standard methods
  - e.g. rational quadratic spline

SU(3) example:



Parallel Transport Convolution Networks

Normal Convolution:

$$\phi(x) \mapsto \sum_{\delta} c_{\delta} \phi(x + \delta)$$

Parallel transport convolution:

Wilson line

$$PTCL[\phi](x) = \sum_{\delta} c_{\delta} W(x, x + \delta)\phi(x + \delta)$$

$$\phi_a(x) \mapsto A[U](x)\phi_a(x) + B[U](x, y)\phi_f(y)$$
$$\phi_f(x) \mapsto \phi_f(x)$$

$$B[U](x, y)\phi_f(y) = PTCL[PTCL[\ldots PTCL[\phi]]]$$

Example: Scalar Field Theory

[Dinh et al, 1605.08803] [Albergo et al., 1904.12072]

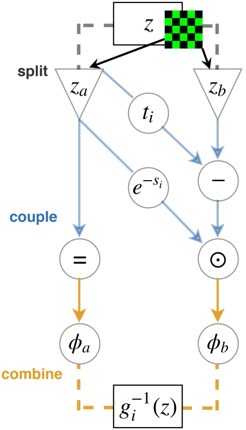
- Fields $\phi(x) \in \mathbb{R}$, target $p(\phi) \propto e^{-S(\phi)}$
- Split $z \to z_a, z_f$ active/frozen
  - Typically: even/odd checkerboard

$$\phi_f = z_f$$
$$\phi_a = e^{s(z_f)} \odot z_a + t(z_f)$$

Arbitrary functions

- Inverse:

$$z_f = \phi_f$$
$$z_a = e^{-s(\phi_f)} \odot (\phi_a - t(\phi_f))$$



- Tractable Jacobian: $\det J = \prod_i e^{s(\phi_f)_i}$
- Compose alternating transforms $(\phi_a, \phi_f) \leftrightarrow (\phi_f, \phi_a)$

Reverse KL Training

- Model density $q(\phi)$, target $p(\phi) = \frac{1}{Z} e^{-S(\phi)}$
- Reverse Kullback Leibler (KL) loss $\mathcal{L}$:

$$\begin{aligned}\mathcal{L} &= D_{KL}(q||p) \\ &= \int \mathrm{d}\phi\, q(\phi) \log \frac{q(\phi)}{p(\phi)} \\ &= \mathbb{E}_{\phi\sim q}\left[\log q(\phi) + S(\phi)\right] + \log Z\end{aligned}$$

Model samples

Constant

($\Rightarrow$ can ignore)

Key facts

$D_{KL}(q||p) \geq 0$

$D_{KL}(q||p) = 0 \Leftrightarrow q = p$