# On impossibility of learning in a reparameterization covariant way

Timothy Holy
Washington University Medical School
holy@pcg.wustl.edu

Ilya Nemenman
KITP, UCSB
nemenman@kitp.ucsb.edu

# Background: Bayesian inference of probability density

# Background: Bayesian inference of probability density

$$Q(x) = \begin{cases} \frac{1}{l_0} \mathrm{e}^{\phi(x)} \\ \phi^2(x) \end{cases}$$ Enforcing positivity of density

# Background: Bayesian inference of probability density

$$Q(x) = \begin{cases} \frac{1}{l_0} \mathrm{e}^{\phi(x)} \\ \phi^2(x) \end{cases}$$

Enforcing positivity of density

$$\mathcal{P}[\phi(x)] \quad = \quad \frac{1}{Z} \exp\left\{ -\frac{\ell^{2\eta-1}}{2} \int dx \left( \frac{\partial^\eta \phi}{\partial x^\eta} \right)^2 \right\} \delta\left[ \int dx\, Q(\phi(x)) - 1 \right]$$

# Background: Bayesian inference of probability density

$$Q(x) = \begin{cases} \frac{1}{l_0} \mathrm{e}^{\phi(x)} \\ \phi^2(x) \end{cases}$$ Enforcing positivity of density

$$\mathcal{P}[\phi(x)] \;=\; \frac{1}{Z} \exp\left\{ -\frac{\ell^{2\eta-1}}{2} \int dx \left( \frac{\partial^\eta \phi}{\partial x^\eta} \right)^2 \right\} \delta\left[ \int dx\, Q(\phi(x)) - 1 \right]$$

Consistent, bias and variance are known.

# Background: Bayesian inference of probability density

$$Q(x) = \begin{cases} \frac{1}{l_0} \mathrm{e}^{\phi(x)} \\ \phi^2(x) \end{cases}$$ Enforcing positivity of density

$$\mathcal{P}[\phi(x)] = \frac{1}{Z} \exp\left\{ -\frac{\ell^{2\eta-1}}{2} \int dx \left( \frac{\partial^\eta \phi}{\partial x^\eta} \right)^2 \right\} \delta\left[ \int dx\, Q(\phi(x)) - 1 \right]$$

Consistent, bias and variance are known.

$$\mathrm{Var}\, \psi(x) \propto (NP(x))^{1/2\eta-1},$$ where $\psi(x) = \phi(x) - \phi_{\mathrm{true}}(x)$

# Background: reparameterization problem

$$x \longrightarrow z = z(x)$$

$$Q(x) \longrightarrow Q(z) = Q(x(z)) \left| \frac{dx}{dz} \right|$$

# Background: reparameterization problem

$$
\begin{aligned}
x &\longrightarrow z = z(x) \\
Q(x) &\longrightarrow Q(z) = Q(x(z)) \left| \frac{dx}{dz} \right|
\end{aligned}
$$

The prior above is not reparameterization–invariant.
Thus reparameterization covariance does not hold.

# Background: reparameterization covariant learning?

$$Q(x) \;=\; \sqrt{|g(x)|}\,\widetilde{Q}(x) = \sqrt{|g|}\widetilde{Q}(\widetilde{\phi}(x))$$

# Background: reparameterization covariant learning?

$$
\begin{aligned}
Q(x) &= \sqrt{|g(x)|}\,\widetilde{Q}(x) = \sqrt{|g|}\widetilde{Q}(\widetilde{\phi}(x)) \\
\mathcal{P}[\widetilde{\phi}(x)] &= \frac{1}{Z}\exp\left\{-\frac{1}{2}\int dx \sqrt{|g|}^{2\eta-1}\left(\frac{\partial^\eta \widetilde{\phi}}{\partial x^\eta}\right)^2\right\} \\
&\times\ \delta\left[\int dx\,\sqrt{|g|}\widetilde{Q}(\phi(x)) - 1\right]
\end{aligned}
$$

# Background: reparameterization covariant learning?

$$
\begin{aligned}
Q(x) &= \sqrt{|g(x)|}\,\widetilde{Q}(x) = \sqrt{|g|}\widetilde{Q}(\widetilde{\phi}(x)) \\
\mathcal{P}[\widetilde{\phi}(x)] &= \frac{1}{Z}\exp\left\{-\frac{1}{2}\int dx\,\sqrt{|g|}^{2\eta-1}\left(\frac{\partial^\eta\widetilde{\phi}}{\partial x^\eta}\right)^2\right\} \\
&\times\ \delta\left[\int dx\,\sqrt{|g|}\widetilde{Q}(\phi(x)) - 1\right]
\end{aligned}
$$

Is this really a solution?

# Background: reparameterization covariant learning?

$$
\begin{aligned}
Q(x) &= \sqrt{|g(x)|}\,\widetilde{Q}(x) = \sqrt{|g|}\widetilde{Q}(\widetilde{\phi}(x)) \\
\mathcal{P}[\widetilde{\phi}(x)] &= \frac{1}{Z}\exp\left\{-\frac{1}{2}\int dx\,\sqrt{|g|}^{2\eta-1}\left(\frac{\partial^\eta\widetilde{\phi}}{\partial x^\eta}\right)^2\right\} \\
&\times\ \delta\left[\int dx\,\sqrt{|g|}\widetilde{Q}(\phi(x)) - 1\right]
\end{aligned}
$$

Is this really a solution?

What is to prevent variability of $g$?

# Suspicion: one dimension

In one dimension

# Suspicion: one dimension

In one dimension

- all differential–geometric properties are due to embedding (parameterization);

# Suspicion: one dimension

In one dimension

- all differential–geometric properties are due to embedding (parameterization);
- no intrinsic curvature to identify complexity;

# Suspicion: one dimension

In one dimension

- all differential–geometric properties are due to embedding (parameterization);
- no intrinsic curvature to identify complexity;
- No way to regularize metric covariantly.

# Counterargument: definitions

Learning operator $L$:

$$L\{x_i,\, i = 1 \dots N\} = Q(x)$$

# Counterargument: definitions

Learning operator $L$:

$$L\{x_i,\, i = 1 \dots N\} = Q(x)$$

Reparameterization operator $R_z$:

$$R_z x \;\; = \;\; z(x)$$

# Counterargument: definitions

Learning operator $L$:

$$L\{x_i,\, i = 1 \ldots N\} = Q(x)$$

Reparameterization operator $R_z$:

$$\begin{aligned} R_z x &= z(x) \\ R_z Q(x) &= Q(x(z)) J(z) \qquad \text{$Q(x)$ is non–singular} \\ & \quad J^{-1}(z) = |dx/dz| \end{aligned}$$

# Counterargument: definitions

Learning operator $L$:

$$L\{x_i,\, i = 1 \ldots N\} = Q(x)$$

Reparameterization operator $R_z$:

$$\begin{aligned} R_z x &= z(x) \\ R_z Q(x) &= Q(x(z)) J(z) \qquad Q(x) \text{ is non–singular} \\ & \quad J^{-1}(z) = |dx/dz| \end{aligned}$$

Reparameterization covariance:

$$[R_z, L] = 0$$

# Counterargument: the essence

Choose reparameterization:

$$z_i = R_z x_i = x_i$$

# Counterargument: the essence

Choose reparameterization:

$$z_i = R_z x_i = x_i$$

Then:

$$LR_z\{x_i\} \;=\; L\{x_i\} \equiv Q(x)$$

# Counterargument: the essence

Choose reparameterization:

$$z_i = R_z x_i = x_i$$

Then:

$$\begin{aligned} L R_z\{x_i\} &= L\{x_i\} \equiv Q(x) \\ R_z L\left\{x_i\right\} &= R_z\, Q(x) = J(z) Q(z) \end{aligned}$$

# Counterargument: the essence

Choose reparameterization:

$$z_i = R_z x_i = x_i$$

Then:

$$\begin{aligned} L R_z \{x_i\} &= L\{x_i\} \equiv Q(x) \\ R_z L\,\{x_i\} &= R_z\, Q(x) = J(z) Q(z) \\ [R_a, L] &= (J-1)L \end{aligned}$$

# Counterexample: result

- $[R_z, L]$ is zero for $z = x$.

# Counterexample: result

- $[R_z, L]$ is zero for $z = x$.
- $[R_z, L]$ is zero for $L\{x_i\} = 1/N \sum \delta(x - x_i)$ (overfits hopelessly).

# Counterexample: result

- $[R_z, L]$ is zero for $z = x$.
- $[R_z, L]$ is zero for $L\{x_i\} = 1/N \sum \delta(x - x_i)$ (overfits hopelessly).
- $[R_z, L]$ nonzero otherwise.

# Counterexample: result

- $[R_z, L]$ is zero for $z = x$.
- $[R_z, L]$ is zero for $L\{x_i\} = 1/N \sum \delta(x - x_i)$ (overfits hopelessly).
- $[R_z, L]$ nonzero otherwise.

Reason: There are infinitely many ways to reparameterize $\{x_i\}$ into equally spaced $\{z_i\}$. Without a priori constraints on coordinates, the data are uninformative.

# Reparameterization problem: generalization, previous history

- Any nontrivially transforming quantity will have the same problem.

# Reparameterization problem: generalization, previous history

- Any nontrivially transforming quantity will have the same problem.
- Cucker and Smale: learning error bounded by the determinant of the operator mapping between assumed measure and the (unknown) true one [equivalently, $J(\mathrm{uniform}, \mathrm{true}) < \infty$].

# Reparameterization problem: generalization, previous history

- Any nontrivially transforming quantity will have the same problem.
- Cucker and Smale: learning error bounded by the determinant of the operator mapping between assumed measure and the (unknown) true one [equivalently, $J(\mathrm{uniform}, \mathrm{true}) < \infty$].
- Learning is minimizing risk:

$$\mathcal{R} = \int dx Q(x) \mathcal{L}(Q, x) \,.$$

# Reparameterization problem: generalization, previous history

- Any nontrivially transforming quantity will have the same problem.
- Cucker and Smale: learning error bounded by the determinant of the operator mapping between assumed measure and the (unknown) true one [equivalently, $J(\mathrm{uniform}, \mathrm{true}) < \infty$].
- Learning is minimizing risk:

$$\mathcal{R} = \int dx Q(x) \mathcal{L}(Q, x) \, .$$

If no constraints on coordinates, then $\exists g(x), \Delta X : \mu(\Delta X) \to 0, R(\Delta X) \to \mathrm{number}$ (or $\infty$).

# Approximate covariance?

$$\mathrm{Var}\,\psi(x) \propto \frac{1}{N^{\alpha} P^{\beta}(x)}$$

# Approximate covariance?

$$\operatorname{Var}\psi(x) \propto \frac{1}{N^{\alpha}P^{\beta}(x)}$$

Bounds can be build through

Chebyshev inequality.

# Approximate covariance?

$$\mathrm{Var}\,\psi(x) \propto \frac{1}{N^{\alpha} P^{\beta}(x)}$$

Bounds can be build through Chebyshev inequality.

- No uniform bounds exist.

# Approximate covariance?

$$\mathrm{Var}\,\psi(x) \propto \frac{1}{N^{\alpha} P^{\beta}(x)}$$

Bounds can be build through Chebyshev inequality.

- No uniform bounds exist.
- Reparameterization may make $P(x) \to 0$ and $\mathrm{Var}\,\psi(x) \to \infty$.

# Approximate covariance?

$$\mathrm{Var}\,\psi(x) \propto \frac{1}{N^\alpha P^\beta(x)}$$

Bounds can be build through Chebyshev inequality.

- No uniform bounds exist.
- Reparameterization may make $P(x) \to 0$ and $\mathrm{Var}\,\psi(x) \to \infty$.
- This is because coordinate system $\Leftrightarrow$ probability density, and smoothness is defined in a particular coordinate system.

# Approximate covariance?

$$\operatorname{Var}\psi(x) \propto \frac{1}{N^{\alpha}P^{\beta}(x)}$$

Bounds can be build through Chebyshev inequality.

- No uniform bounds exist.
- Reparameterization may make $P(x) \to 0$ and $\operatorname{Var}\psi(x) \to \infty$.
- This is because coordinate system $\Leftrightarrow$ probability density, and smoothness is defined in a particular coordinate system.

Even approximate covariance does not hold if arbitrary transformations are allowed.

# Approximate covariance: assumptions needed

If $P(x) \geq P_0 > 0$ (equivalently, uniform measure is absolutely continuous with respect to the true measure), then

# Approximate covariance: assumptions needed

If $P(x) \geq P_0 > 0$ (equivalently, uniform measure is absolutely continuous with respect to the true measure), then

$$\operatorname{Var} \psi(x) \propto \frac{1}{N^{\alpha} P_0^{\beta}} .$$

# Approximate covariance: assumptions needed

If $P(x) \geq P_0 > 0$ (equivalently, uniform measure is absolutely continuous with respect to the true measure), then

$$\operatorname{Var}\psi(x) \propto \frac{1}{N^\alpha P_0^\beta}\,.$$

- Assuming "reasonable" coordinate system leads to uniform bounds and approximate covariance for some class of coordinates.

# Approximate covariance: assumptions needed

If $P(x) \geq P_0 > 0$ (equivalently, uniform measure is absolutely continuous with respect to the true measure), then

$$\mathrm{Var}\,\psi(x) \propto \frac{1}{N^{\alpha} P_0^{\beta}}\,.$$

- Assuming "reasonable" coordinate system leads to uniform bounds and approximate covariance for some class of coordinates.
- Assumption must not be hard, but may be smoothly enforced by priors, e. g.:

$$\mathcal{P}[\phi] \propto \exp\left[\lambda \int dx \log Q\right]$$

# Approximate covariance: assumptions needed

If $P(x) \geq P_0 > 0$ (equivalently, uniform measure is absolutely continuous with respect to the true measure), then

$$\mathrm{Var}\,\psi(x) \propto \frac{1}{N^\alpha P_0^\beta}\,.$$

- Assuming "reasonable" coordinate system leads to uniform bounds and approximate covariance for some class of coordinates.
- Assumption must not be hard, but may be smoothly enforced by priors, e. g.:

$$\mathcal{P}[\phi] \propto \exp\left[\lambda \int dx \log Q\right]$$ (choosing $\lambda$?)

# Covariance–approximation tradeoff

As $P_0 \longrightarrow 0$ with uniform bound still finite, full covariance is restored.

# Covariance–approximation tradeoff

As $P_0 \to 0$ with uniform bound still finite, full covariance is restored. But:

$$\mathrm{Var}\,\psi(x)\,P_0^{\beta} \propto \frac{1}{N^{\alpha}}\,.$$

# Covariance–approximation tradeoff

As $P_0 \to 0$ with uniform bound still finite, full covariance is restored. But:

$$\operatorname{Var}\psi(x)\, P_0^{\beta} \propto \frac{1}{N^{\alpha}}\,.$$

Thus there is a tradeoff between the quality of covariance (as measured by $P_0$) and the approximation (as measured by $\operatorname{Var}\psi$).

# Covariance–approximation tradeoff

As $P_0 \to 0$ with uniform bound still finite, full covariance is restored. But:

$$\operatorname{Var}\psi(x)\, P_0^{\beta} \propto \frac{1}{N^{\alpha}}\,.$$

Thus there is a tradeoff between the quality of covariance (as measured by $P_0$) and the approximation (as measured by $\operatorname{Var}\psi$).

- Balance is governed by $N$.

# Covariance–approximation tradeoff

As $P_0 \to 0$ with uniform bound still finite, full covariance is restored. But:

$$\mathrm{Var}\,\psi(x)\,P_0^\beta \propto \frac{1}{N^\alpha}\,.$$

Thus there is a tradeoff between the quality of covariance (as measured by $P_0$) and the approximation (as measured by $\mathrm{Var}\,\psi$).

- Balance is governed by $N$.
- Details of the balance are assumption–dependent.

# Covariance–approximation tradeoff

As $P_0 \to 0$ with uniform bound still finite, full covariance is restored. But:

$$\mathrm{Var}\,\psi(x)\,P_0^{\beta} \propto \frac{1}{N^{\alpha}}\,.$$

Thus there is a tradeoff between the quality of covariance (as measured by $P_0$) and the approximation (as measured by $\mathrm{Var}\,\psi$).

- Balance is governed by $N$.
- Details of the balance are assumption–dependent.
- We conjecture such tradeoff to be a general feature.

# Covariance–approximation tradeoff

As $P_0 \to 0$ with uniform bound still finite, full covariance is restored. But:

$$\mathrm{Var}\,\psi(x)\,P_0^{\beta} \propto \frac{1}{N^{\alpha}}\,.$$

Thus there is a tradeoff between the quality of covariance (as measured by $P_0$) and the approximation (as measured by $\mathrm{Var}\,\psi$).

- Balance is governed by $N$.
- Details of the balance are assumption–dependent.
- We conjecture such tradeoff to be a general feature.
- How can this balance be self–consistently selected?

# Implications

- The world seems to be continuous.

# Implications

- The world seems to be continuous.
- Various convergence bounds are usually proven for finite alphabets, pre-defined partitionings (structures), finite–parameter systems.

# Implications

- The world seems to be continuous.
- Various convergence bounds are usually proven for finite alphabets, pre-defined partitionings (structures), finite–parameter systems.
- One should be careful that chosen quantization is appropriate.

# Implications

- The world seems to be continuous.
- Various convergence bounds are usually proven for finite alphabets, pre-defined partitionings (structures), finite–parameter systems.
- One should be careful that chosen quantization is appropriate.
- One should check if the obtained "great learning performance" is a result of constraining parameterization and/or discretization.