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

\[ Q(x) = \begin{cases} \frac{1}{l_0} e^{\phi(x)} & \text{Enforcing positivity of density} \\ \phi^2(x) \end{cases} \]
Background: Bayesian inference of probability density

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

Enforcing positivity of density

\[ P[\phi(x)] = \frac{1}{Z} \exp \left\{ -\frac{\ell^2 \eta - 1}{2} \int dx \left( \frac{\partial \eta \phi}{\partial x} \right)^2 \right\} \delta \left[ \int dx Q(\phi(x)) - 1 \right] \]
Background: Bayesian inference of probability density

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

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

Consistent, bias and variance are known.
Background: Bayesian inference of probability density

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

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.

\[ \text{Var} \psi(x) \propto (NP(x))^{1/2\eta^{-1}}, \text{ where } \psi(x) = \phi(x) - \phi_{\text{true}}(x) \]
Background: reparameterization problem

\[ x \rightarrow z = z(x) \]

\[ Q(x) \rightarrow Q(z) = Q(x(z)) \left| \frac{dx}{dz} \right| \]
Background: reparameterization problem

\[ x \rightarrow z = z(x) \]

\[ Q(x) \rightarrow Q(z) = Q(x(z)) \left| \frac{dx}{dz} \right| \]

The prior above is not reparameterization–invariant.

Thus reparameterization covariance does not hold.
Background: reparameterization covariant learning?

\[ Q(x) = \sqrt{|g(x)|} \tilde{Q}(x) = \sqrt{|g|}\tilde{Q}(\tilde{\phi}(x)) \]
Background: reparameterization covariant learning?

\[ Q(x) = \sqrt{|g(x)|} \tilde{Q}(x) = \sqrt{|g|} \tilde{Q}(\tilde{\phi}(x)) \]

\[ \mathcal{P}[\tilde{\phi}(x)] = \frac{1}{Z} \exp \left\{ -\frac{1}{2} \int dx \sqrt{|g|}^{2\eta-1} \left( \frac{\partial \eta \tilde{\phi}}{\partial x^\eta} \right)^2 \right\} \]

\[ \times \delta \left[ \int dx \sqrt{|g|} \tilde{Q}(\phi(x)) - 1 \right] \]
Background: reparameterization covariant learning?

\[ Q(x) = \sqrt{|g(x)|} \tilde{Q}(x) = \sqrt{|g|}\tilde{Q}(\tilde{\phi}(x)) \]

\[ \mathcal{P}[\tilde{\phi}(x)] = \frac{1}{Z} \exp \left\{ -\frac{1}{2} \int dx \sqrt{|g|}^{2\eta-1} \left( \frac{\partial \eta \tilde{\phi}}{\partial x^\eta} \right)^2 \right\} \]

\[ \times \delta \left[ \int dx \sqrt{|g|}\tilde{Q}(\phi(x)) - 1 \right] \]

Is this really a solution?

Ilya Nemenman, Negative results workshop, NIPS’02, December 14, 2002
Background: reparameterization covariant learning?

\[ Q(x) = \sqrt{|g(x)|} \quad \tilde{Q}(x) = \sqrt{|g|} \tilde{Q}(\phi(x)) \]

\[ \mathcal{P}[\tilde{\phi}(x)] = \frac{1}{Z} \exp \left\{ -\frac{1}{2} \int dx \sqrt{|g|^{2\eta-1}} \left( \frac{\partial^n \tilde{\phi}}{\partial x^n} \right)^2 \right\} \]

\[ \times \delta \left[ \int dx \sqrt{|g|} \tilde{Q}(\phi(x)) - 1 \right] \]

Is this really a solution? What is to prevent variability of \( g \)?

Ilya Nemenman, Negative results workshop, NIPS’02, December 14, 2002
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 $L$:

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

Learning operator $L$:

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

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 $R_z$:

$$R_z x = z(x)$$

$$R_z Q(x) = Q(x(z)) J(z) \quad Q(x) \text{ is non-singular}$$

$$J^{-1}(z) = |dx/dz|$$
Counterargument: definitions

Learning operator $L$:

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

Reparameterization operator $R_z$:

$$R_z x = z(x)$$

$$R_z Q(x) = Q(x(z)) J(z) \quad Q(x) \text{ is non–singular}$$

$$J^{-1}(z) = \left| \frac{dx}{dz} \right|$$

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:

\[ LR_z \{ x_i \} = L \{ x_i \} \equiv Q(x) \]

\[ R_z L \{ x_i \} = R_z Q(x) = J(z)Q(z) \]
Counterargument: the essence

Choose reparameterization:

\[ z_i = R_z x_i = x_i \]

Then:

\[ LR_z \{ x_i \} = L \{ x_i \} \equiv Q(x) \]
\[ R_z L \{ x_i \} = R_z Q(x) = J(z)Q(z) \]
\[ [R_a, L] = (J - 1)L \]
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 \(L\{x_i\} = \frac{1}{N} \sum \delta(x - x_i)\) (overfits hopelessly).
Counterexample: result

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

- \([R_z, L]\) is zero for \(z = x\).
- \([R_z, L]\) is zero for \(L\{x_i\} = \frac{1}{N} \sum \delta (x - x_i)\) (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 \( J(\text{uniform, true}) < \infty \).
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 \([\text{equivalently, } J(\text{uniform}, \text{true}) < \infty]\).
- 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 \([\text{equivalently, } J(\text{uniform, true}) < \infty]\).
- Learning is minimizing risk:

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

If no constraints on coordinates, then
\[ \exists g(x), \Delta X : \mu(\Delta X) \to 0, R(\Delta X) \to \text{number (or } \infty). \]
Approximate covariance?

$$\text{Var } \psi(x) \propto \frac{1}{N^\alpha P^\beta(x)}$$
Approximate covariance?

\[ \text{Var } \psi(x) \propto \frac{1}{N^\alpha P^\beta(x)} \]

Bounds can be built through Chebyshev inequality.
Approximate covariance?

\[ \text{Var} \psi(x) \propto \frac{1}{N^\alpha P^\beta(x)} \]

- No uniform bounds exist.

Bounds can be built through Chebyshev inequality.
Approximate covariance?

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

Bounds can be build through Chebyshev inequality.

- No uniform bounds exist.
- Reparameterization may make $P(x) \to 0$ and $\text{Var } \psi(x) \to \infty$. 
Approximate covariance?

\[ \text{Var } \psi(x) \propto \frac{1}{N^\alpha P^\beta(x)} \]

Bounds can be build through Chebyshev inequality.

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

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

Bounds can be build through Chebyshev inequality.

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

\[
\text{Var } \psi(x) \propto \frac{1}{N^\alpha P_0^\beta}.
\]
Approximate covariance: assumptions needed

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

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

- 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

\[
\text{Var } \psi(x) \propto \frac{1}{N^{\alpha} P_0^\beta}.
\]

- 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.:

\[
P[\phi] \propto \exp \left[ \lambda \int dx \log Q \right]
\]
Approximate covariance: assumptions needed

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

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

- 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 $P_0 \to 0$ 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:

$$\text{Var } \psi(x) P_0^\beta \propto \frac{1}{N^\alpha}.$$
Covariance–approximation tradeoff

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

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

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

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

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

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

- Balance is governed by $N$. 

Ilya Nemenman, Negative results workshop, NIPS'02, December 14, 2002
Covariance–approximation tradeoff

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

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

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

- 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:

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

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

- 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:

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

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

- 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.