Transfer Learning Beyond Bounded Density Ratios

Alkis Kalavasis

Ilias Zadik

Manolis Zampetakis

Classical Learning

We observe data (x, y), where $x \sim P$ and $\mathbb{E}[y \quad x] = f(x)$

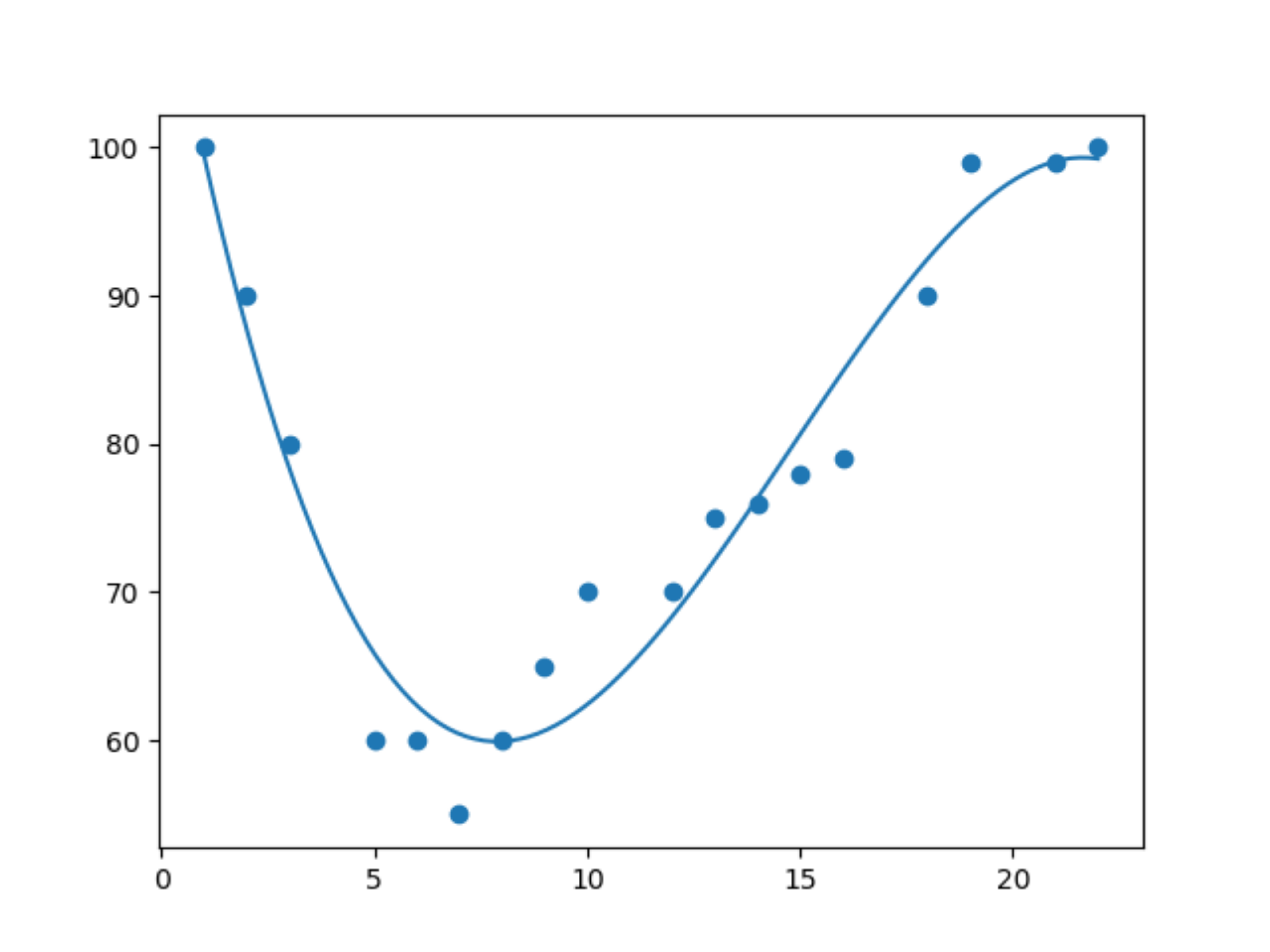
Goal: Find $\hat{f}$ that minimizes

$$\mathrm{err}_P(\hat{f}) \triangleq \mathbb{E}_{x \sim P}\left[(f(x) - \hat{f}(x))^2\right]$$

Classical Learning

We observe d

Goal: Find $\hat{f}$



Transfer Learning

We observe data (x, y), where $x \sim P$ and $\mathbb{E}[y \;\; x] = f(x)$

Goal: Find $\hat{f}$ that minimizes

$$\mathrm{err}_Q(\hat{f}) \triangleq \mathbb{E}_{x \sim Q}\left[(f(x) - \hat{f}(x))^2\right]$$

Transfer Learning

We observe data (x, y), where $x \sim P$ and $\mathbb{E}[y \quad x] = f(x)$

Goal: Find $\hat{f}$ that minimizes

$$\mathrm{err}_Q(\hat{f}) \triangleq \mathbb{E}_{x \sim Q}\left[(f(x) - \hat{f}(x))^2\right]$$

Transfer Learning

We **want to** minimize

$\mathrm{err}_Q(\hat{f})$

vs

We **can** minimize

$\mathrm{err}_P(\hat{f})$

Change of Measure

$$\mathbb{E}_{x\sim Q}\left[(f(x)-\hat{f}(x))^2\right] = \mathbb{E}_{x\sim P}\left[\frac{dQ}{dP}(x)\cdot(f(x)-\hat{f}(x))^2\right]$$

$$\leq \left\|\frac{dQ}{dP}\right\|_\infty \cdot \mathbb{E}_{x\sim P}\left[(f(x)-\hat{f}(x))^2\right]$$

Change of Measure

$$\mathbb{E}_{x\sim Q}\left[(f(x)-\hat{f}(x))^2\right] = \mathbb{E}_{x\sim P}\left[\frac{dQ}{dP}(x)\cdot(f(x)-\hat{f}(x))^2\right]$$

$$\leq \left\| \frac{dQ}{dP} \right\|_\infty \cdot \mathbb{E}_{x\sim P}\left[(f(x)-\hat{f}(x))^2\right]$$

This makes sense only when $Q \ll P$

Transfer Learning

$$\beta = \left\| \frac{dQ}{dP} \right\|_\infty$$

We **want to** minimize

$$\mathrm{err}_Q(\hat{f})$$

vs

We **can** minimize

$$\beta \cdot \mathrm{err}_P(\hat{f})$$

Transfer Learning

$$\beta = \left\| \frac{dQ}{dP} \right\|_{\infty}$$

We **want to** minimize

$$\mathrm{err}_Q(\hat{f})$$

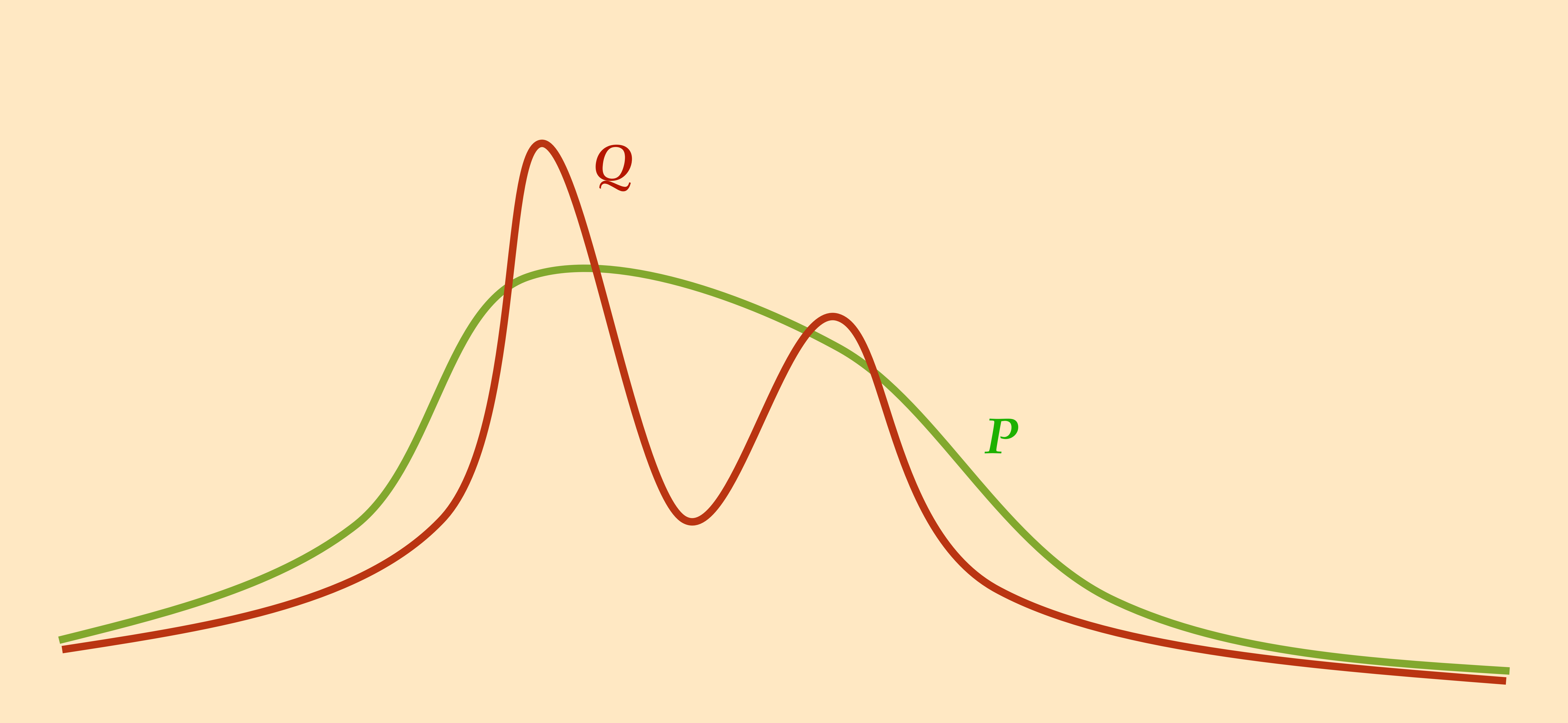
vs

We **can** minimize

$$\beta \cdot \mathrm{err}_P(\hat{f})$$

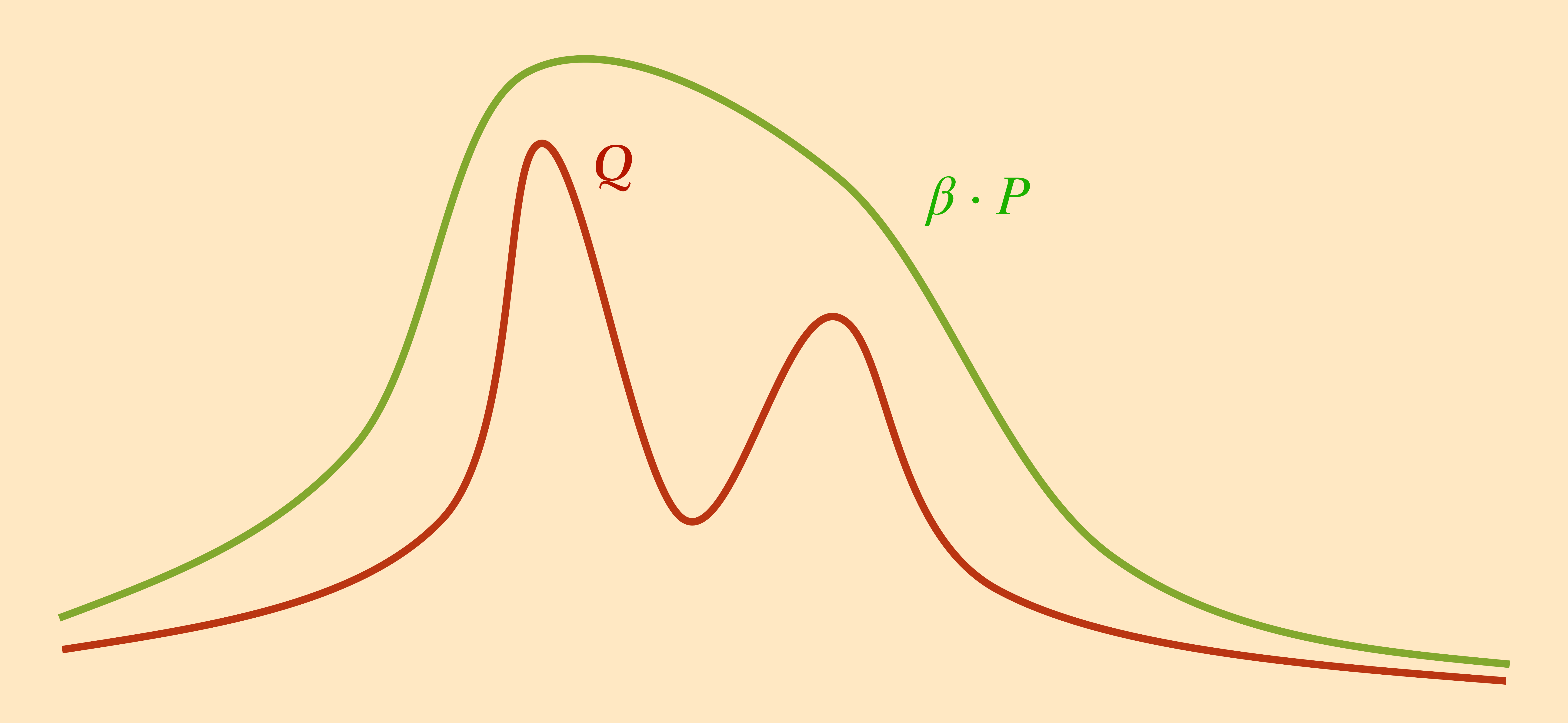
[SSK12, SK12, SKM07, Kpo17, QB13, KM18, CMM10, MPW23, PMW22]: Assume $\beta < \infty$

Transfer Learning

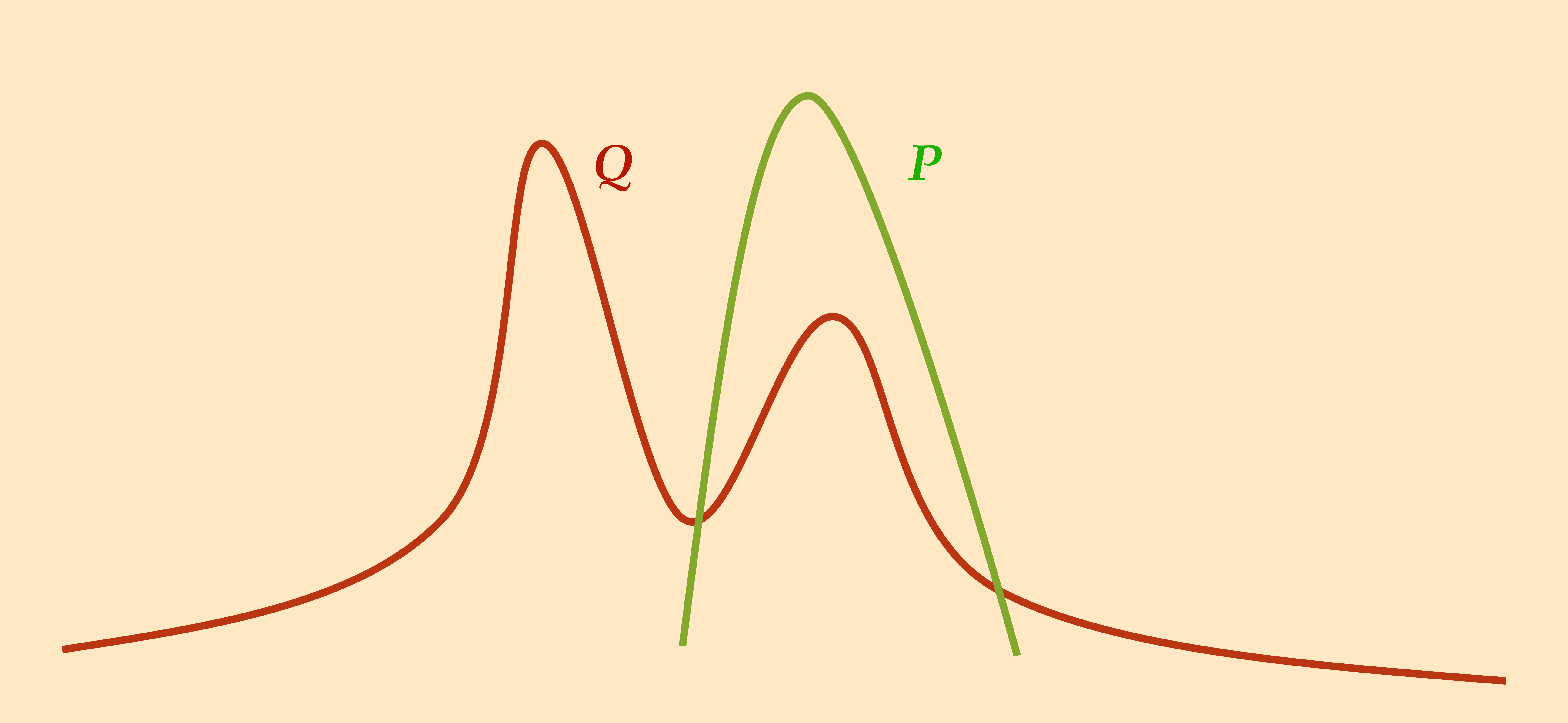


Transfer Learning

$$\beta = \left\| \frac{dQ}{dP} \right\|_\infty$$

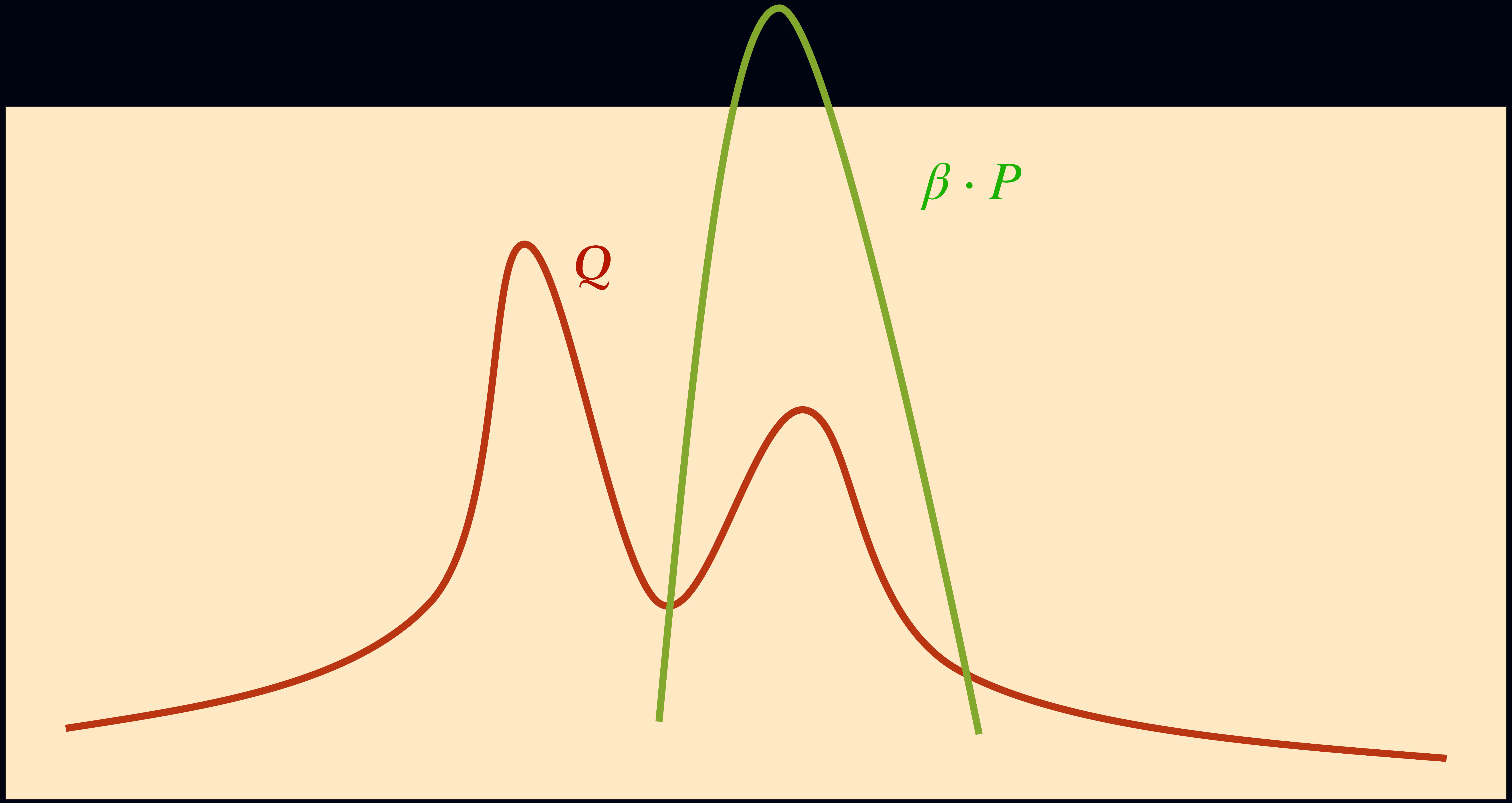


Failure I: Truncation



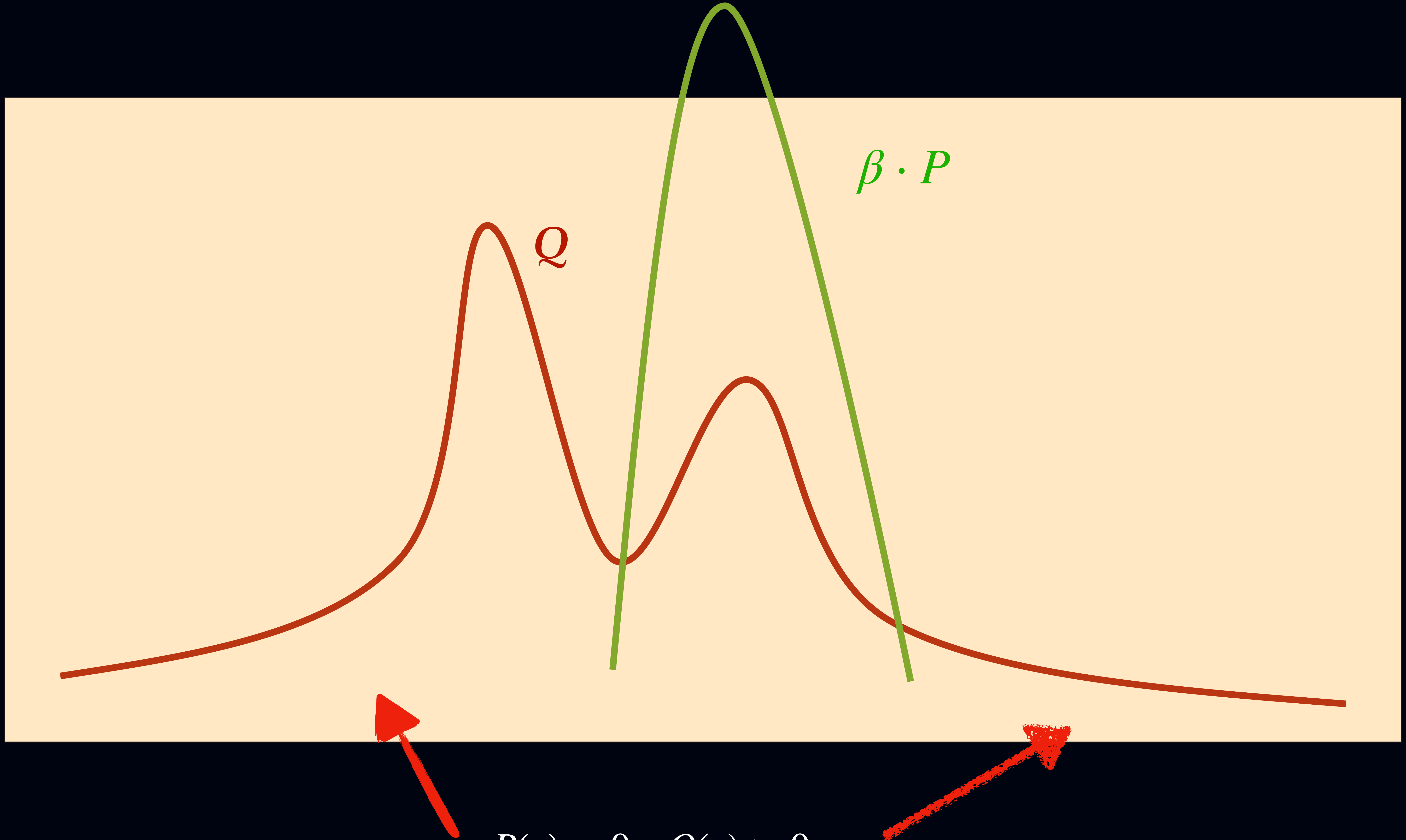
Failure I: Truncation

$$\beta = \left\| \frac{dQ}{dP} \right\|_\infty = \infty$$



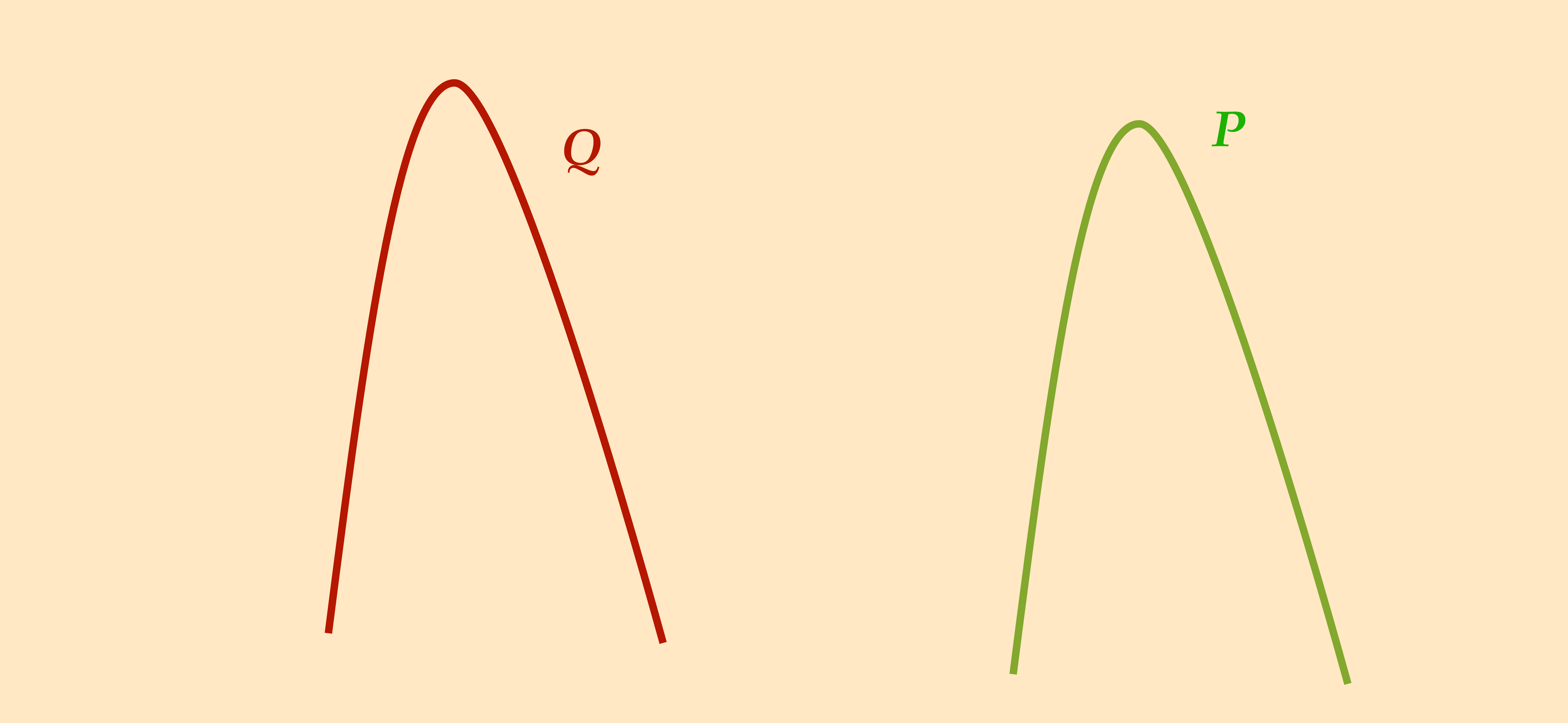
Failure I: Truncation

$$\beta = \left\| \frac{dQ}{dP} \right\|_\infty = \infty$$



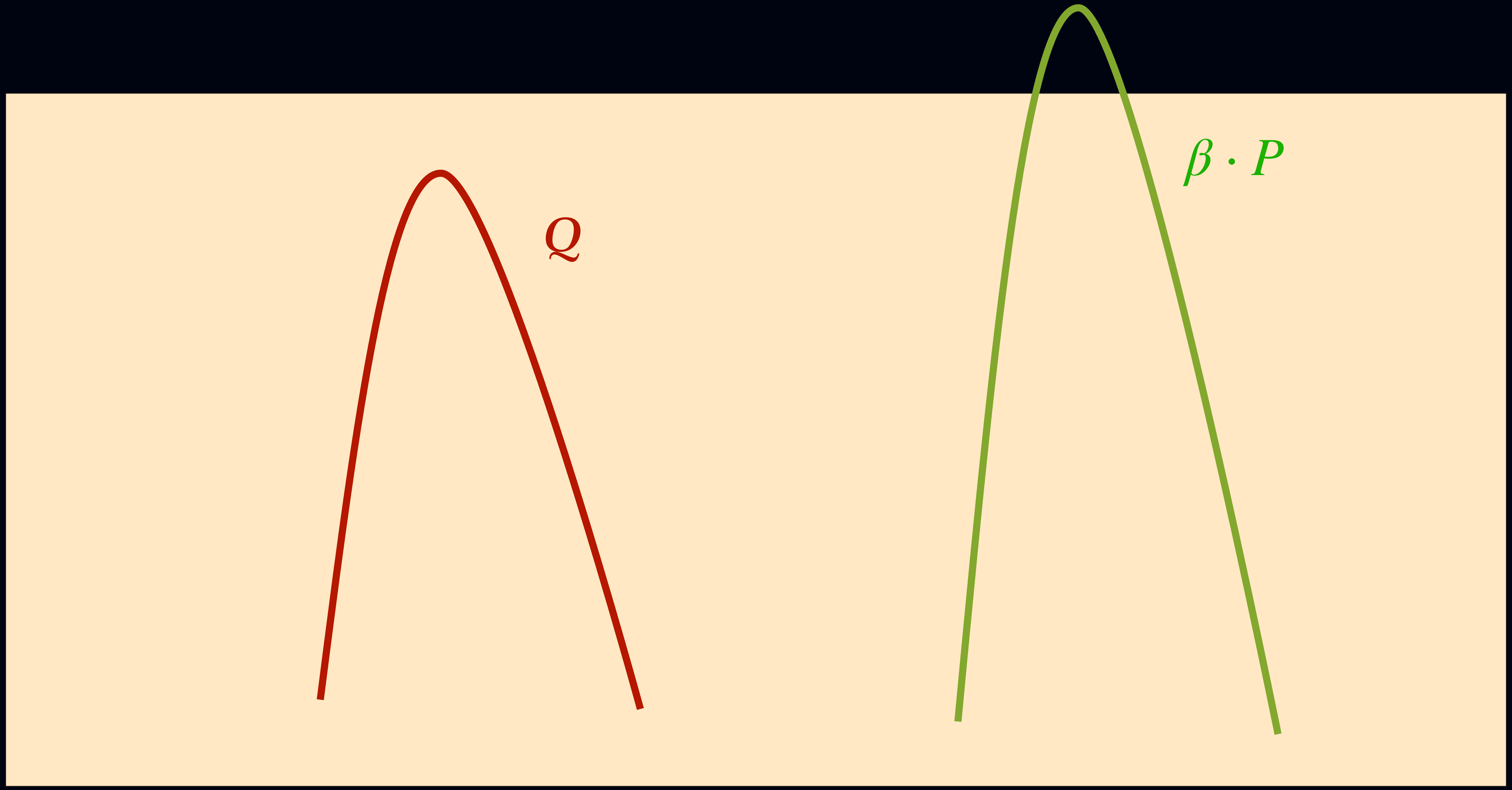
$P(x) = 0, \quad Q(x) > 0$

Failure II: Shift



Failure II: Shift

$$\beta = \left\| \frac{dQ}{dP} \right\|_\infty = \infty$$



Transfer Learning

Observation.

1. Truncated Statistics [DGTZ18, KTZ19, Ple20, NP20, DKTZ21,…]

2. Some classification settings [KM18, HK19]

3. Linear regression with distribution shift [LHL21, GTF+23, ZBGS22, WZB+22]

There are cases where $\left\| \frac{dQ}{dP} \right\|_r \to \infty$ but transfer is possible

$\mu \propto \exp(-\text{convex})$

Our Result

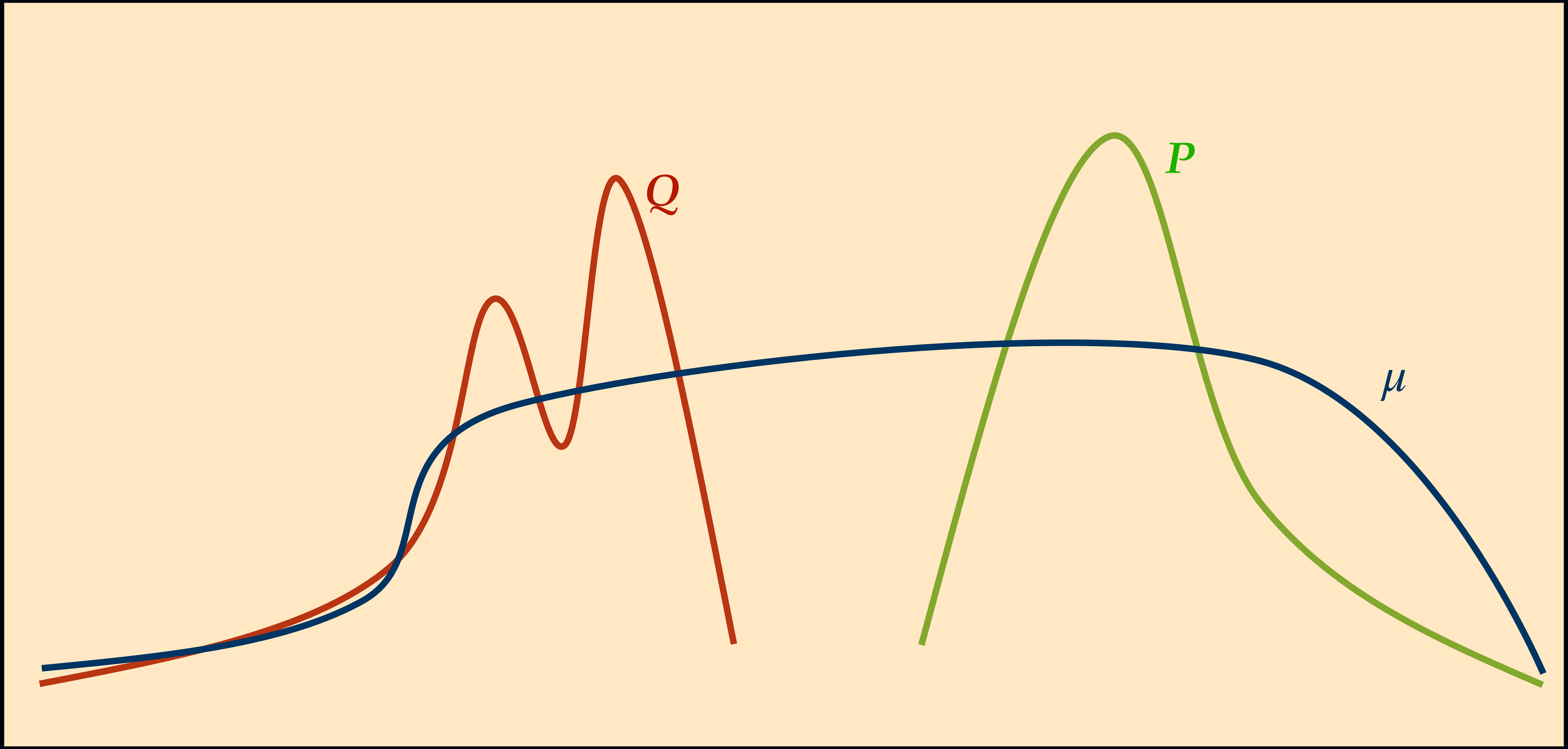
Theorem [**K,** Zadik, Zampetakis '24]

Let f and $\hat{f}$ be degree-k polynomials and μ a log-concave measure. Then:

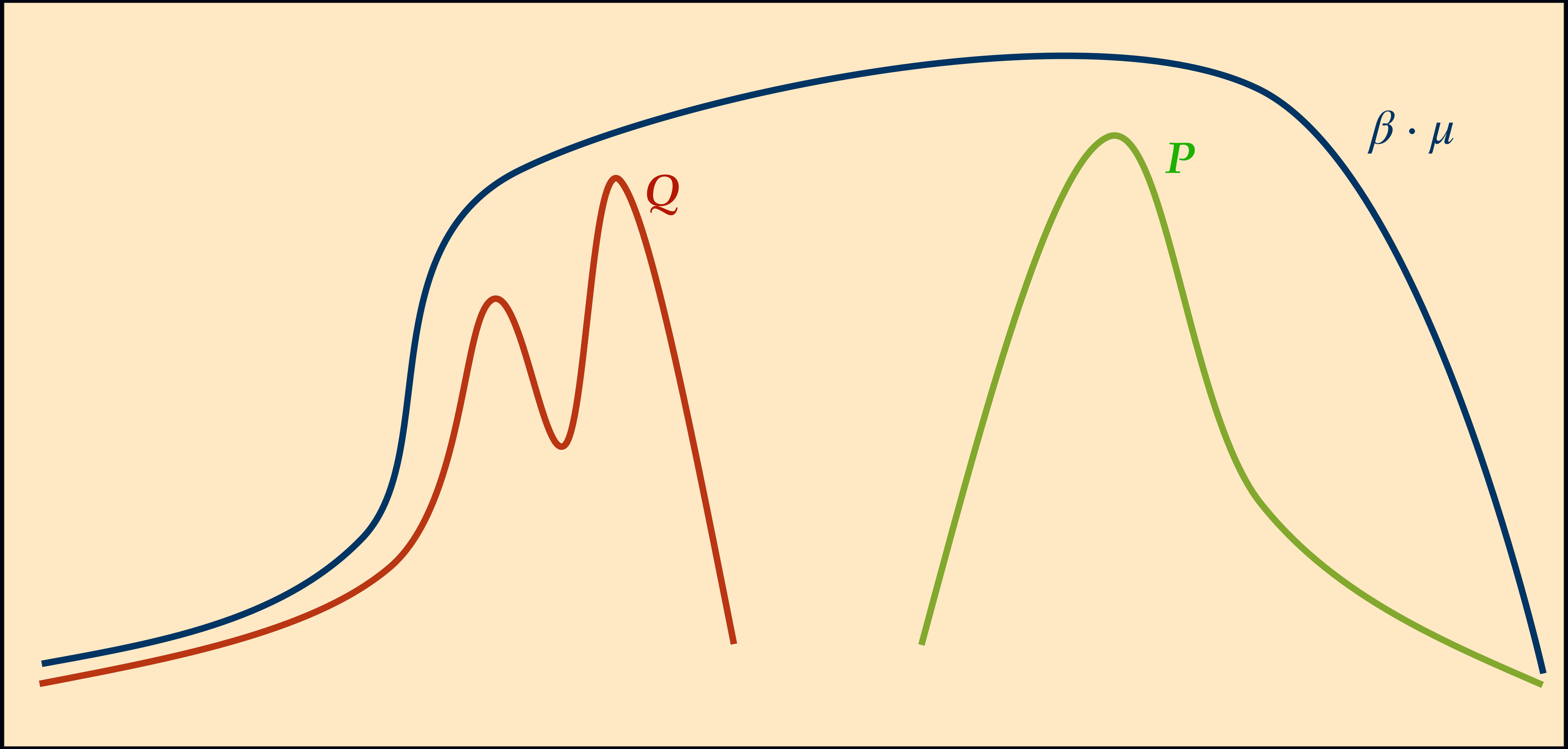
$$\text{err}_Q(\hat{f}) \leq h(k) \cdot \left\| \frac{dQ}{d\mu} \right\|_\infty \cdot \left\| \frac{dP}{d\mu} \right\|_\infty^k \cdot \text{err}_P(\hat{f})$$

$h(k) \leq k^k$

Intuition



Intuition



Our Result

Theorem [**K,** Zadik, Zampetakis '24]

Let f and $\hat{f}$ be degree-k polynomials and μ a log-concave measure. Then:

$$\mathrm{err}_Q(\hat{f}) \leq h(k) \cdot \boxed{\left\| \frac{dQ}{d\mu} \right\|_\infty \cdot \left\| \frac{dP}{d\mu} \right\|_\infty^k} \cdot \mathrm{err}_P(\hat{f})$$

new measure of divergence
sufficient for transferability of polynomials

Comparison with Change of Measure

Theorem [**K,** Zadik, Zampetakis '24]

Let f and $\hat{f}$ be degree-k polynomials and Q a log-concave measure. Then:

$$\mathrm{err}_Q(\hat{f}) \leq h(k) \cdot \left\| \frac{dP}{dQ} \right\|_\infty^k \cdot \mathrm{err}_P(\hat{f})$$

inverse density ratio

Our Result

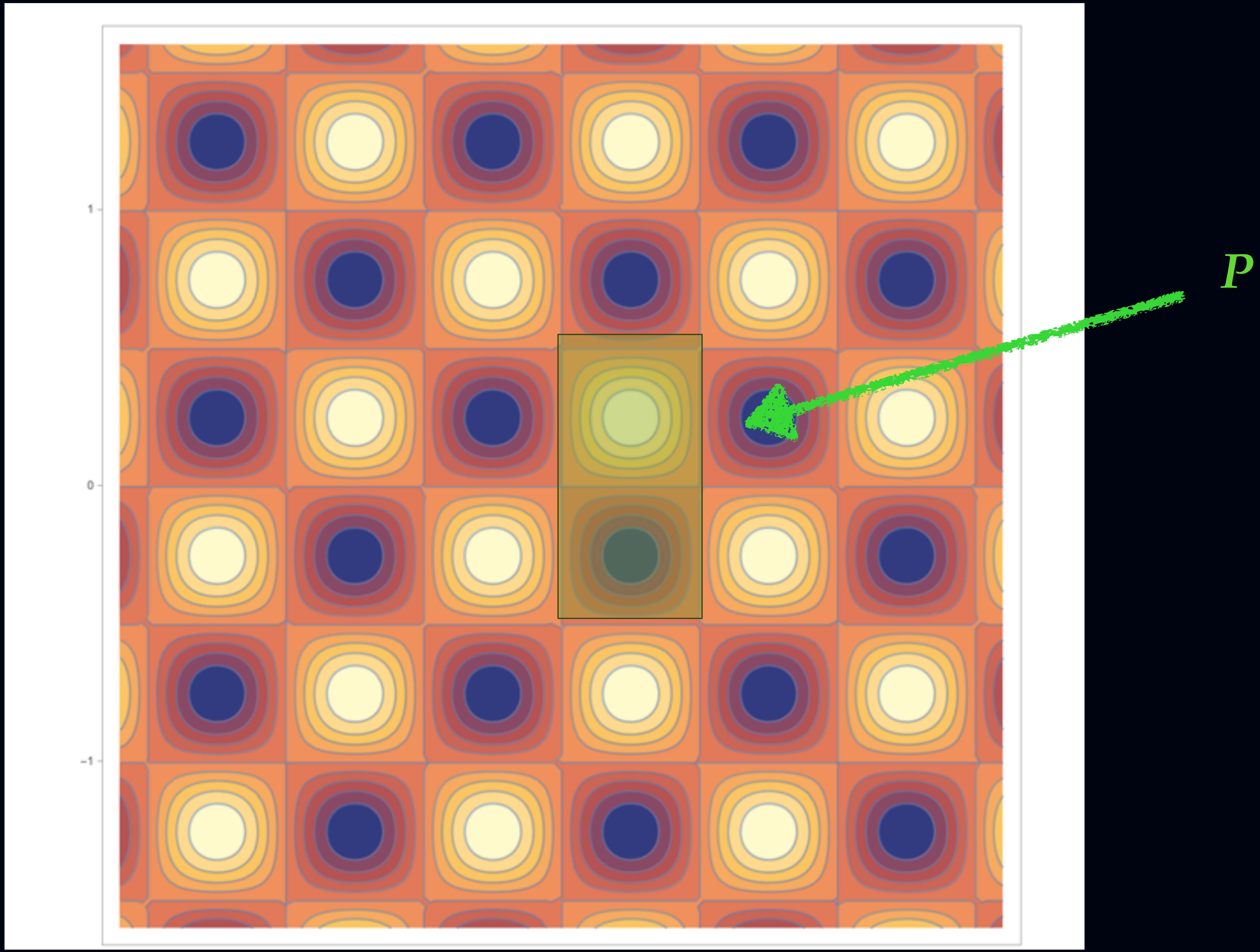
Theorem [**K,** Zadik, Zampetakis ’24]

Let f and $\hat{f}$ be degree-k polynomials and μ a log-concave measure. Then:

$$\mathrm{err}_Q(\hat{f}) \leq h(k) \cdot \left\| \frac{dQ}{d\mu} \right\|_\infty \cdot \left\| \frac{dP}{d\mu} \right\|_\infty^k \cdot \mathrm{err}_P(\hat{f})$$

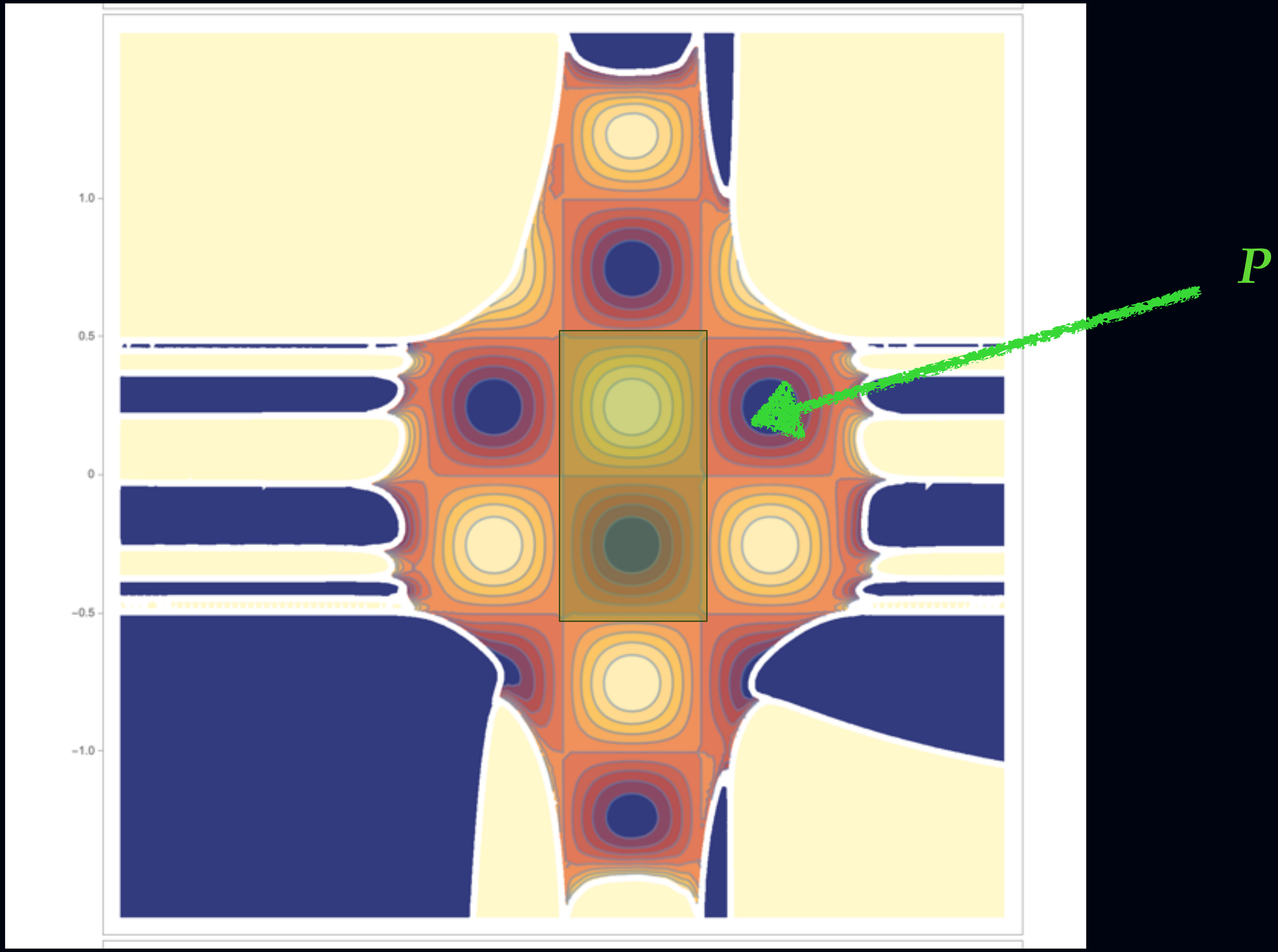
Can we have a similar result for neural networks?

Example: Target Function



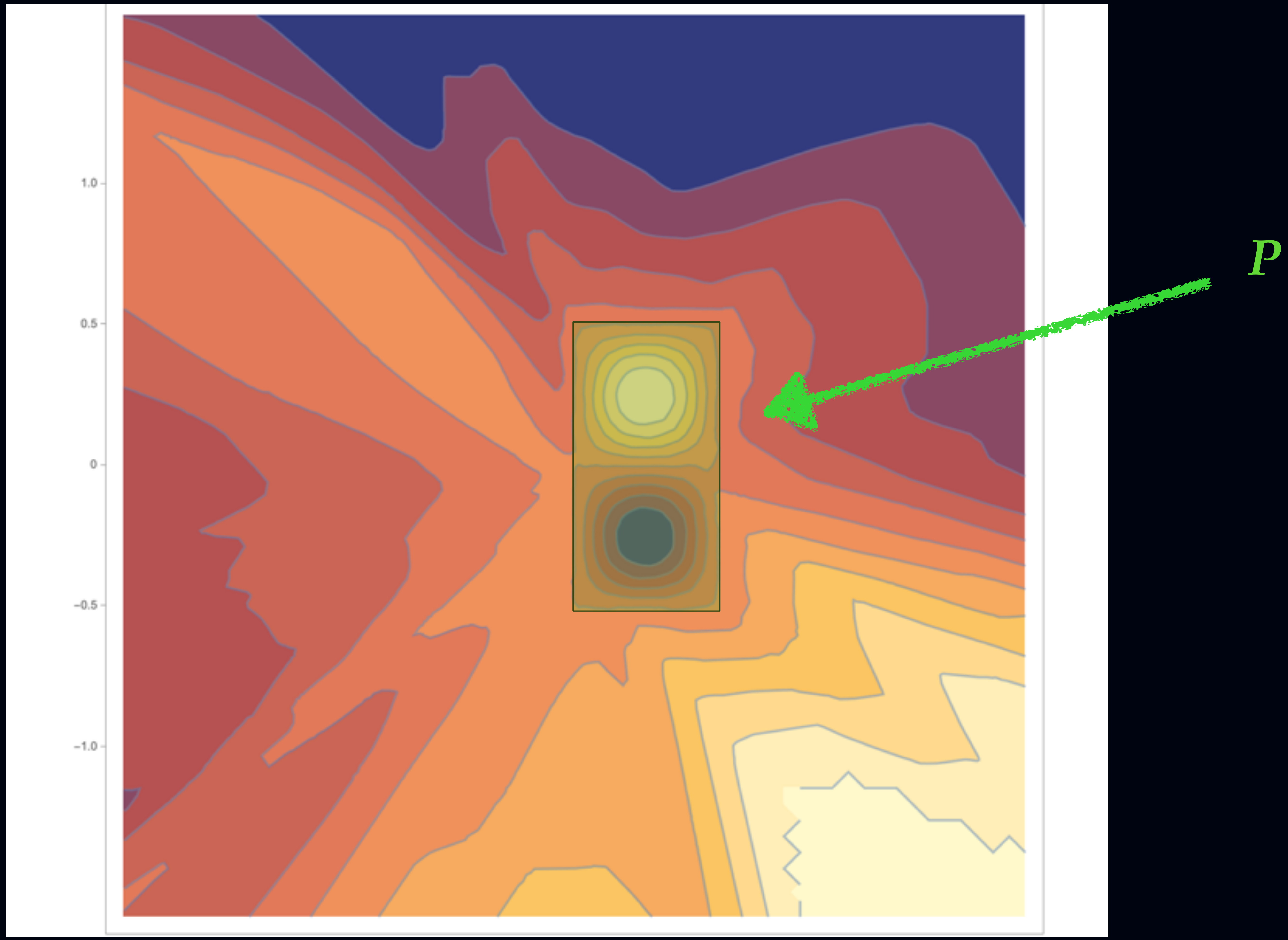
$f : \mathbb{R}^2 \to \mathbb{R}$, not a polynomial

Example: Polynomial Estimator



$\hat{f} : \mathbb{R}^2 \to \mathbb{R}$, polynomial estimator

Example: Neural Networks



$\hat{f} : \mathbb{R}^2 \to \mathbb{R}$, NN estimator trained from P with SGD

Our Result

Theorem [**K,** Zadik, Zampetakis '24]

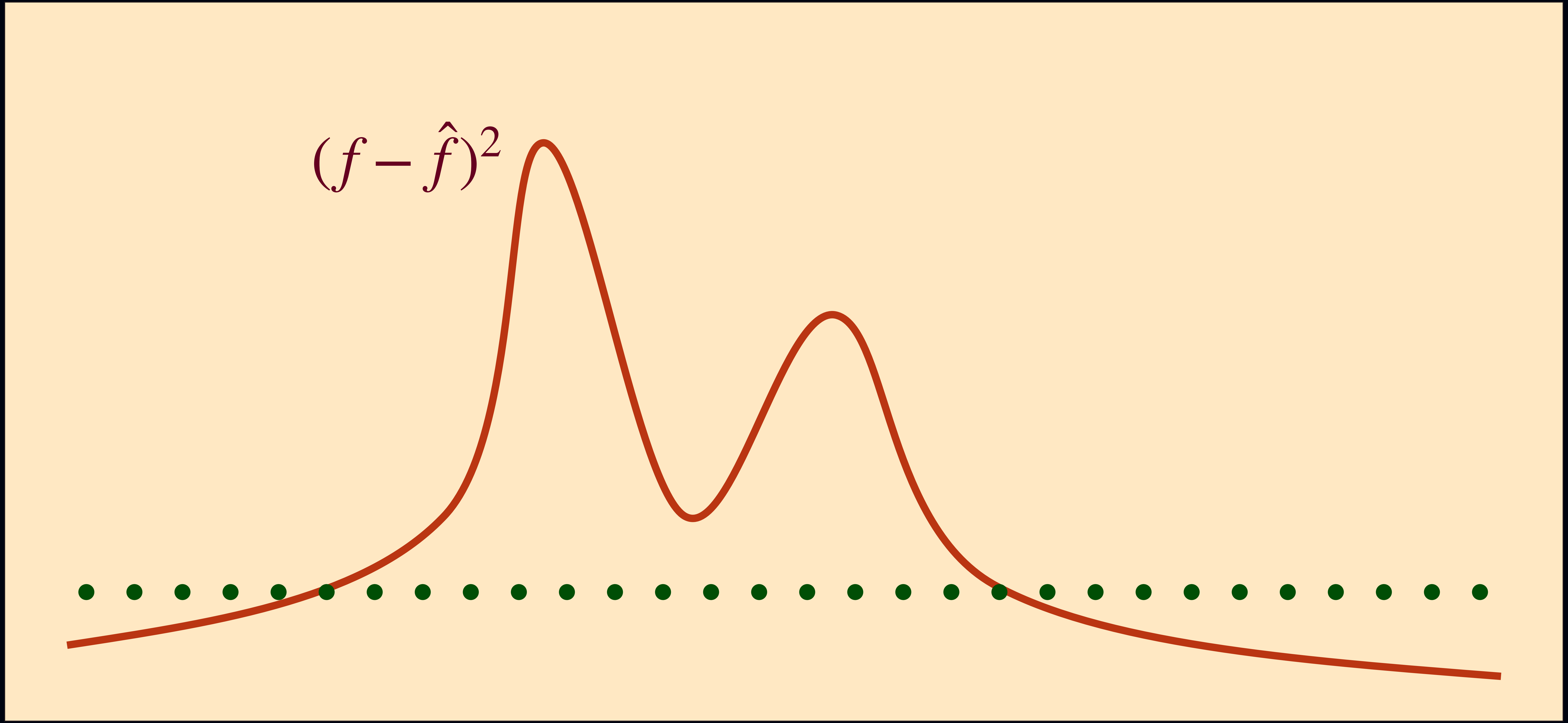
Let f and $\hat{f}$ be degree-k polynomials and μ a log-concave measure. Then:

$$\mathrm{err}_Q(\hat{f}) \le h(k) \cdot \left\| \frac{dQ}{d\mu} \right\|_\infty \cdot \left\| \frac{dP}{d\mu} \right\|_\infty^k \cdot \mathrm{err}_P(\hat{f})$$

Polynomials seem to transfer better than NNs

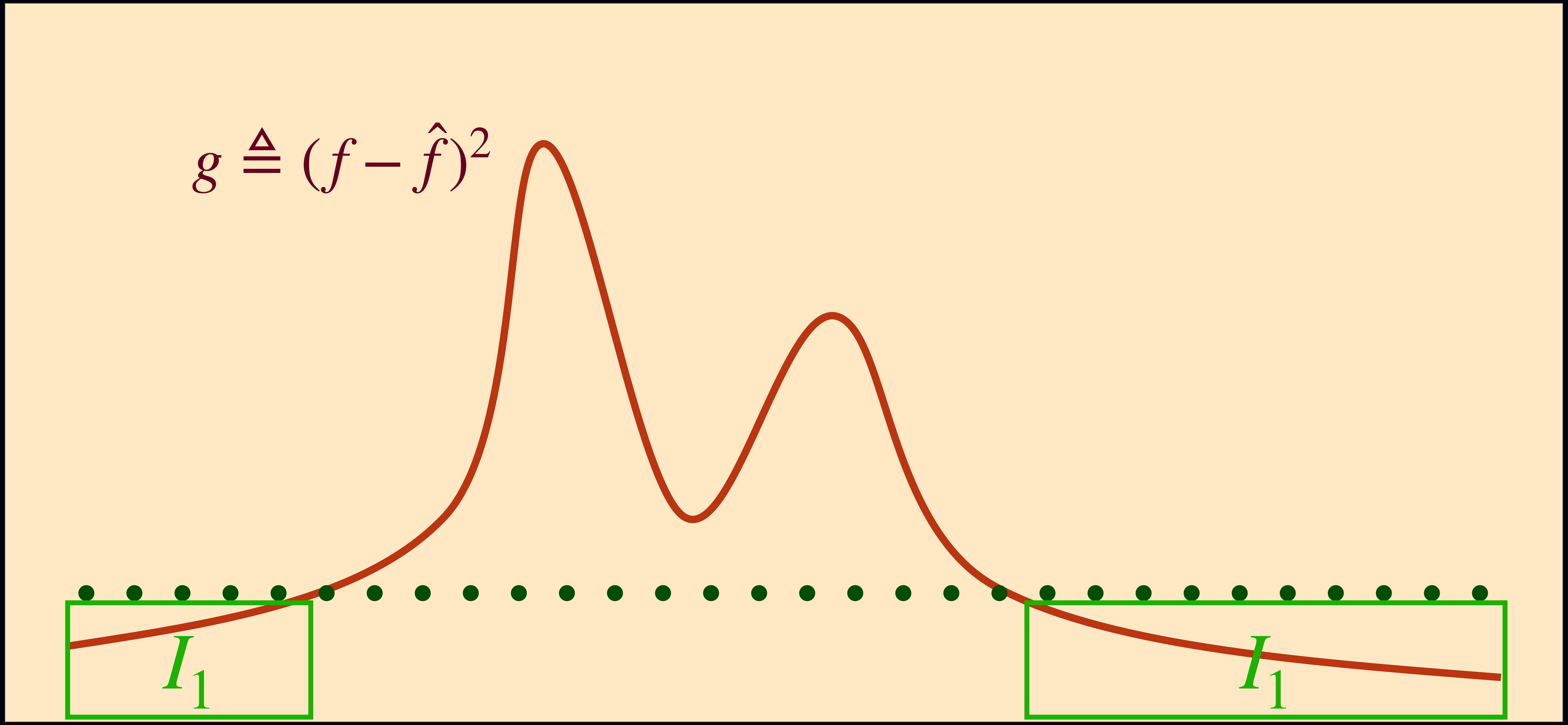
Proof Idea

💡 Anti-concentration implies Extrapolation



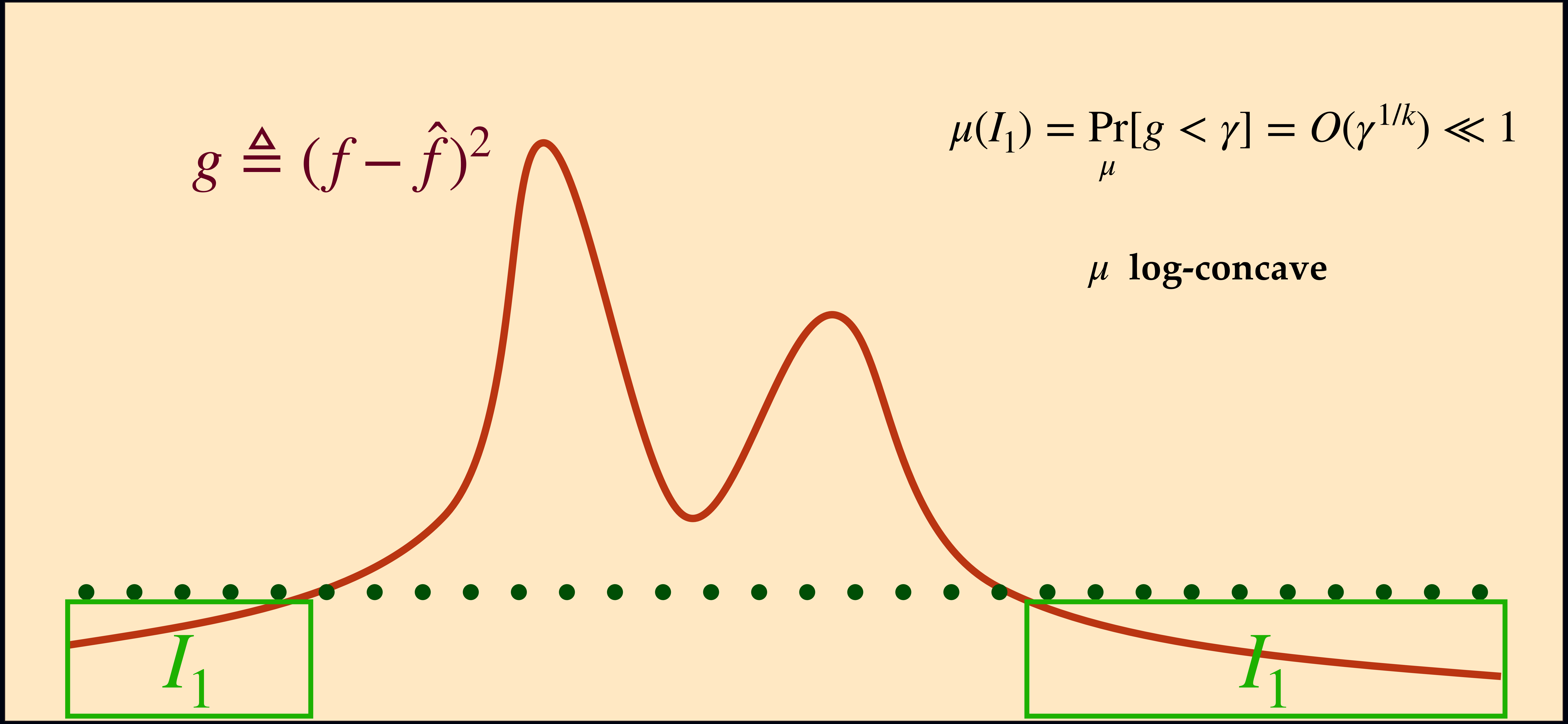
Proof Idea

Anti-concentration implies Extrapolation



Proof Idea

Anti-concentration implies Extrapolation

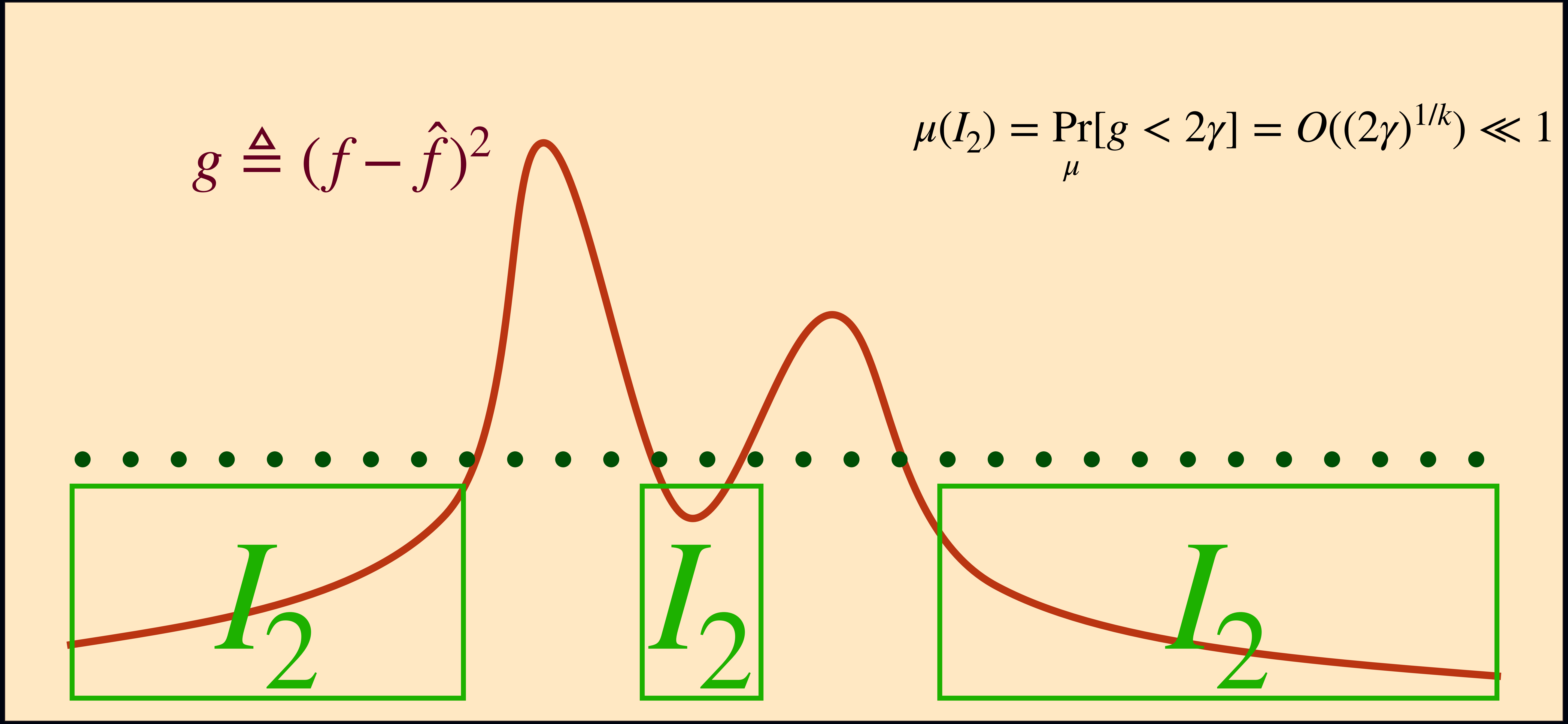


$\mu(I_1) = \Pr_{\mu}[g < \gamma] = O(\gamma^{1/k}) \ll 1$

μ **log-concave**

Proof Idea

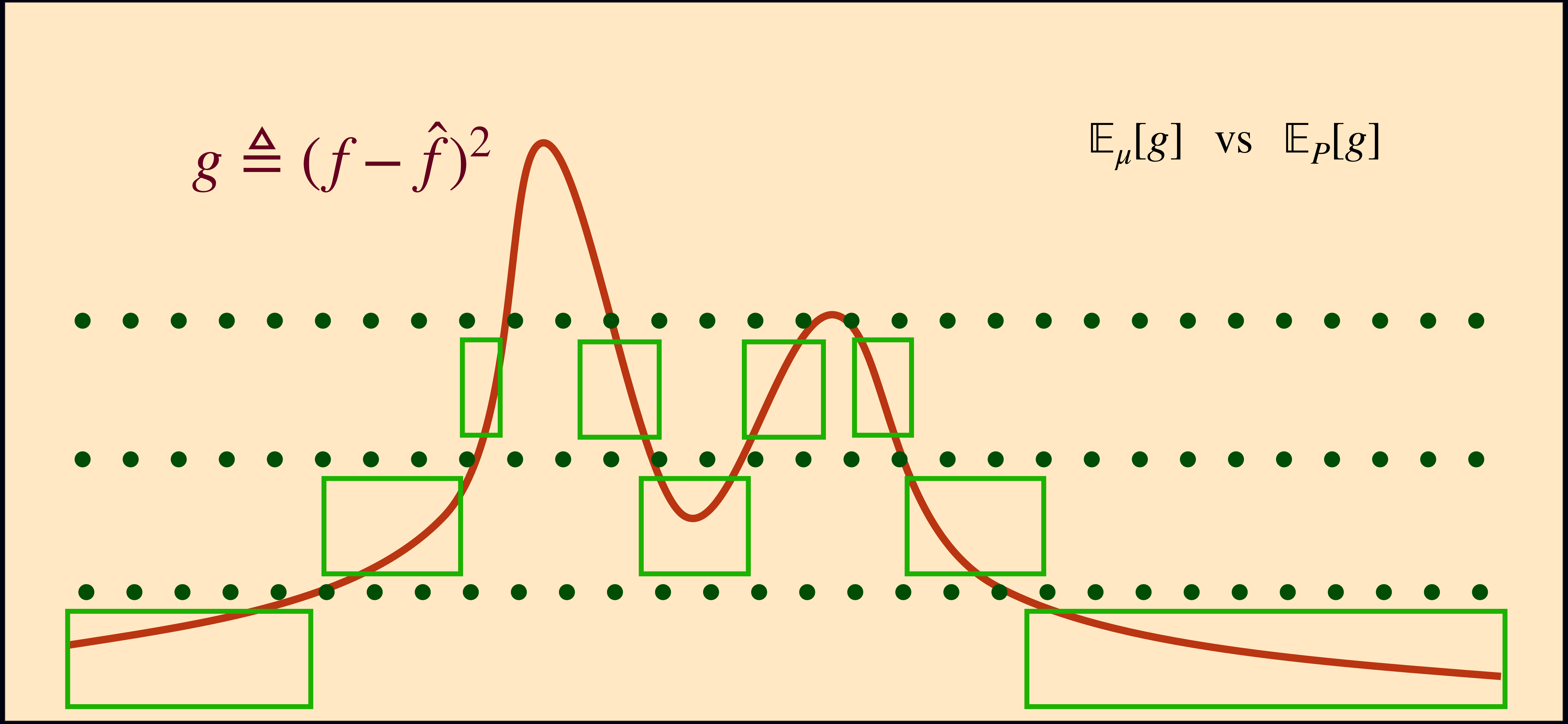
💡 **Anti-concentration implies Extrapolation**



$g \triangleq (f - \hat{f})^2$
$\mu(I_2) = \Pr_{\mu}[g < 2\gamma] = O((2\gamma)^{1/k}) \ll 1$
I_2
I_2
I_2

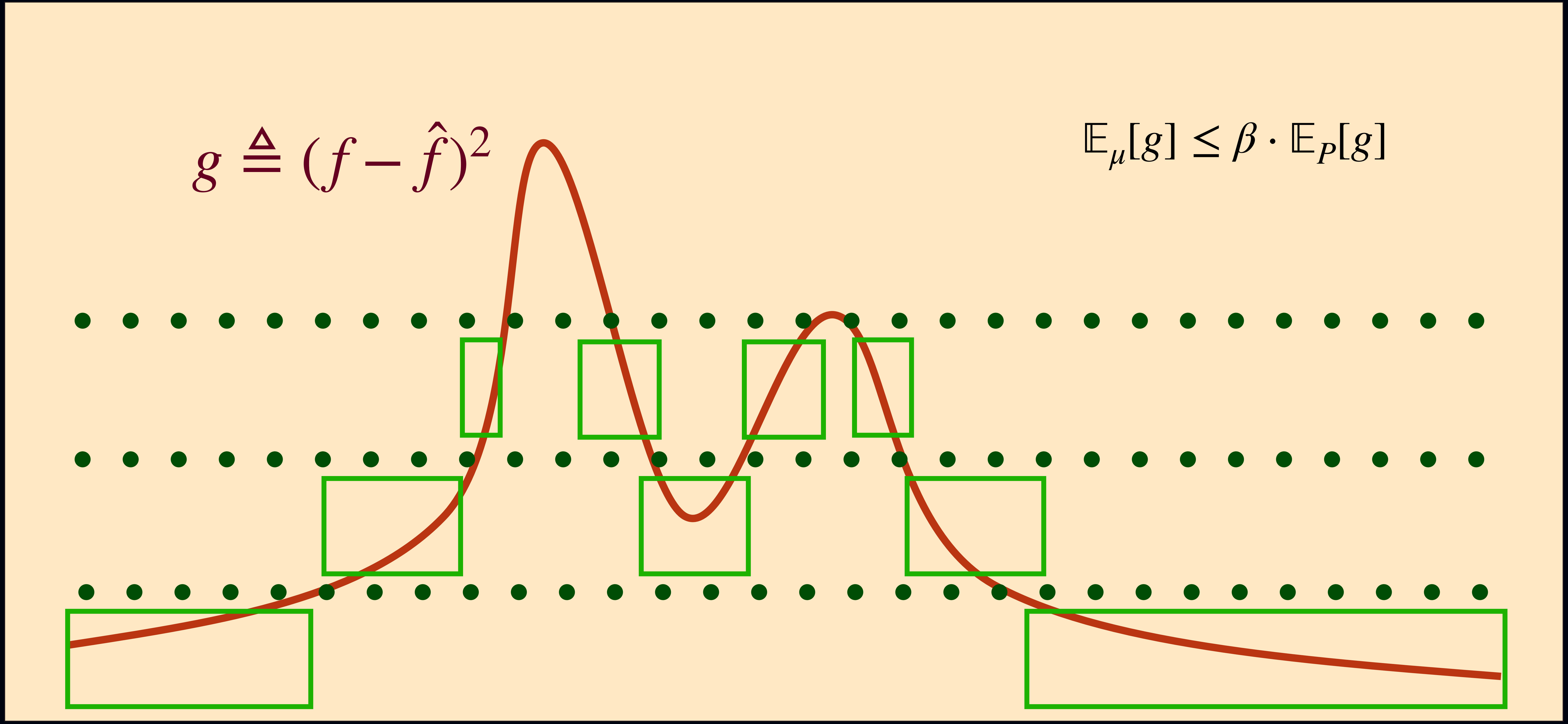
Proof Idea

Anti-concentration implies Extrapolation


$g \triangleq (f - \hat{f})^2$
$\mathbb{E}_\mu[g]$ vs $\mathbb{E}_P[g]$

Proof Idea

Anti-concentration implies Extrapolation



$g \triangleq (f - \hat{f})^2$

$\mathbb{E}_\mu[g] \leq \beta \cdot \mathbb{E}_P[g]$

Comparison with Linear Regression (I)

Is polynomial regression with distribution shift hard?

“Just learn n^k coefficients and transfer without bounded ratios”

Vandermonde matrix in high-dimensions is poorly understood
How to bound the condition number?

Comparison with Linear Regression (II)

$$f_\theta(x) = \theta^\top x$$

$$\mathrm{err}_P(\hat{\theta}) = (\theta - \hat{\theta})^\top \, \mathbb{E}_P[X^\top X] \, (\theta - \hat{\theta})$$

$$\mathrm{err}_Q(\hat{\theta}) = (\theta - \hat{\theta})^\top \, \mathbb{E}_Q[X^\top X] \, (\theta - \hat{\theta})$$

Comparison with Linear Regression (II)

$\Sigma_P = \mathbb{E}_P[xx^\top]$

$f_\theta(x) = \theta^\top x$

$\mathrm{err}_P(\hat{\theta}) = (\theta - \hat{\theta})^\top \ \mathbb{E}_P[X^\top X] \ (\theta - \hat{\theta})$

$\mathrm{err}_Q(\hat{\theta}) = (\theta - \hat{\theta})^\top \ \mathbb{E}_Q[X^\top X] \ (\theta - \hat{\theta})$

Transfer is "related" to $\Sigma_Q \Sigma_P^{-1}$

Rigorous for **specific** estimators in **specific** settings [LHL21, GTF+23]

Comparison with Linear Regression (II)

$f_\theta(x) = \theta^\top x$

$\mathrm{err}_P(\hat{\theta}) = (\theta -$

$\mathrm{err}_Q(\hat{\theta}) = (\theta -$

Transfer is “re

Rigorous for **specific** estimators in **specific** settings [LHL21, GTF+23]

How to control the transfer cost in general?

$$\mathrm{err}_Q(\hat{\theta}) \leq \frac{\lambda_{\max}(\Sigma_Q)}{\lambda_{\min}(\Sigma_P)} \cdot \mathrm{err}_P(\hat{\theta})$$

Our Result

Theorem [K, Zadik, Zampetakis ’24]

Let f and $\hat{f}$ be degree-k polynomials and μ a log-concave measure. Then:

$$\mathrm{err}_Q(\hat{f}) \leq h(k) \cdot \left\| \frac{dQ}{d\mu} \right\|_\infty \cdot \left\| \frac{dP}{d\mu} \right\|_\infty^k \cdot \mathrm{err}_P(\hat{f})$$

+ Arbitrary polynomials
+ Intuitive, not algebraic
+ Extends to Boolean domains
- Needs log-concave bridge

Future Work

1. **Extensions to classification**

2. **Transferability is a property of**

    a. **Model Class**

    b. ***P*, *Q***

    c. **Training Algorithm (Which algorithms could help transfer?)**

3. **Transfer Learning in Other Domains (Adaptive Environments)**

Future Work

1. Extensions to classification

2. Transferability is a property of

    a. Model Class

    b. *P, Q*

    c. Training Algorithm (Which algorithms could help transfer?)

3. Transfer Learning in Other Domains (Adaptive Environments)

Thank You!