Taylor Residual Estimators via Automatic Differentiation

Andrew C. Miller, Nicholas J. Foti, and Ryan P. Adams

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Expectation Objectives

$$X \sim \pi\,, \quad X \in \mathbb{R}^D$$

Random Variable

$$\begin{aligned}\mu_f &\triangleq \mathbb{E}_\pi[f(X)] \\ &= \int f(x)\pi(dx)\end{aligned}$$

Estimand: expectation

Estimate μ_f efficiently

Monte Carlo Estimators

$$x^{(n)} \sim \pi\,, \text{ for } n = 1, \ldots, N$$

$$\hat{\mu}_f = \frac{1}{N} \sum_{n=1}^{N} f(x^{(n)})$$ Monte Carlo Estimator

$$\mathbb{E}\left[\hat{\mu}_f\right] = \mu_f$$ unbiased

$$\mathbb{V}[\hat{\mu}_f]$$ Monte Carlo variance

Expectation Objectives: Examples

e.g. Variational Inference Objective (ELBO)

$p(X, \mathcal{D})\,,\, p(X \mid \mathcal{D})$ model, posterior

$X \sim q_{\boldsymbol{\lambda}}$ posterior approximation

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{X \sim q_{\boldsymbol{\lambda}}}\left[\ln p(X, \mathcal{D}) - \ln q_{\boldsymbol{\lambda}}(X)\right]$$

- Variational Inference
- Importance Sampling
- Entropy Estimation
- Adversarial Learning

What if we have additional information?

differentiable structure in $f(x)$

$$\frac{\partial f}{\partial x}, \frac{\partial^2 f}{\partial x^2}, \dots$$

computable moments of $\pi(x)$

(distributions with moment generating functions)

$$\mathcal{M}_{x_0}^{(m)} = \mathbb{E}_{X \sim \pi}\left[(X - x_0)^m\right]$$

$$= \int (x - x_0)^m \pi(dx)$$

What if we have additional information?

differentiable structure in $f(x)$ computable moments of $\pi(x)$

expand

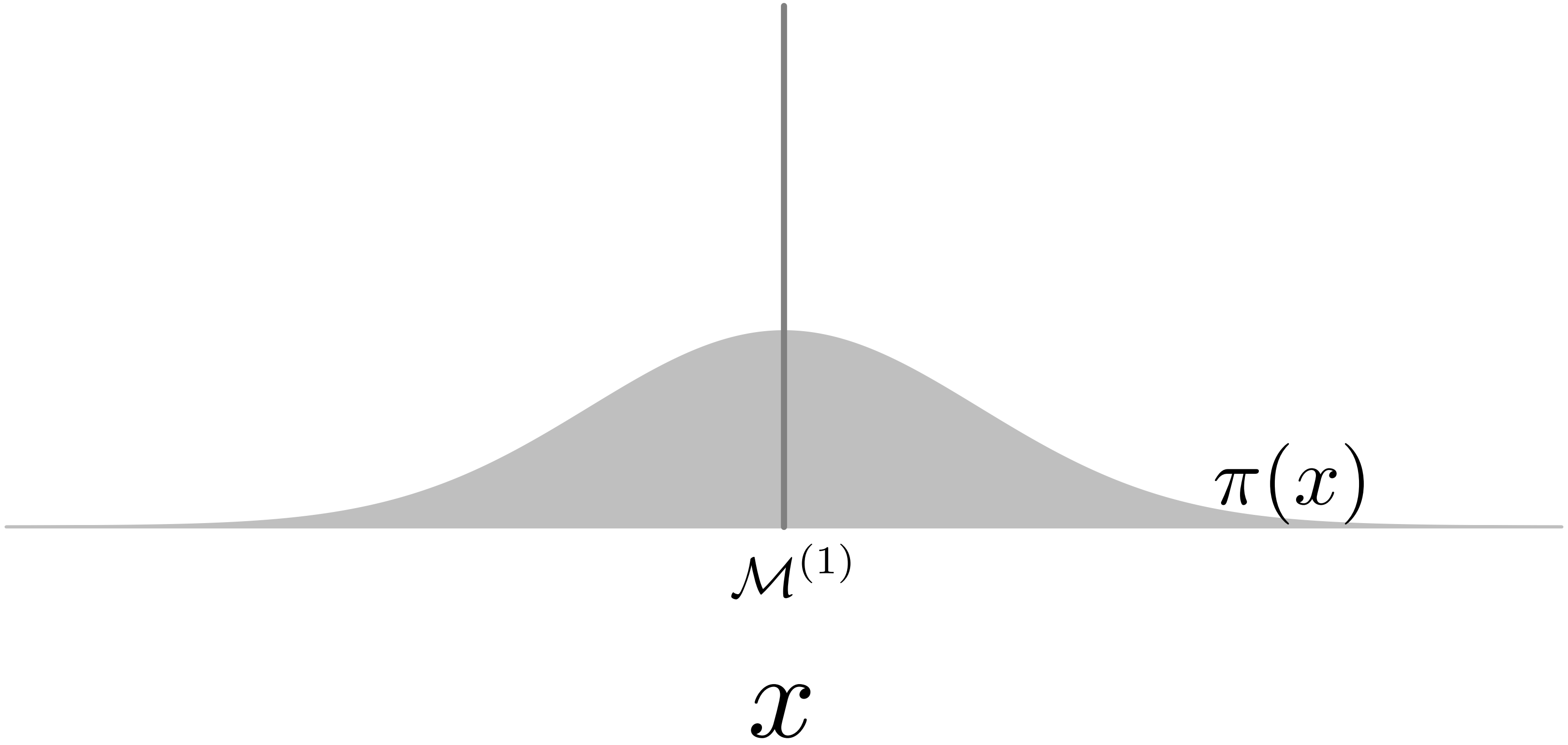
$$f(x) = \underbrace{f(x_0) + (x - x_0)^\intercal \frac{\partial f}{\partial x}(x_0)}_{\text{Taylor}} + \underbrace{R_{x_0}(x)}_{\text{residual}}$$

integrate out lower moments

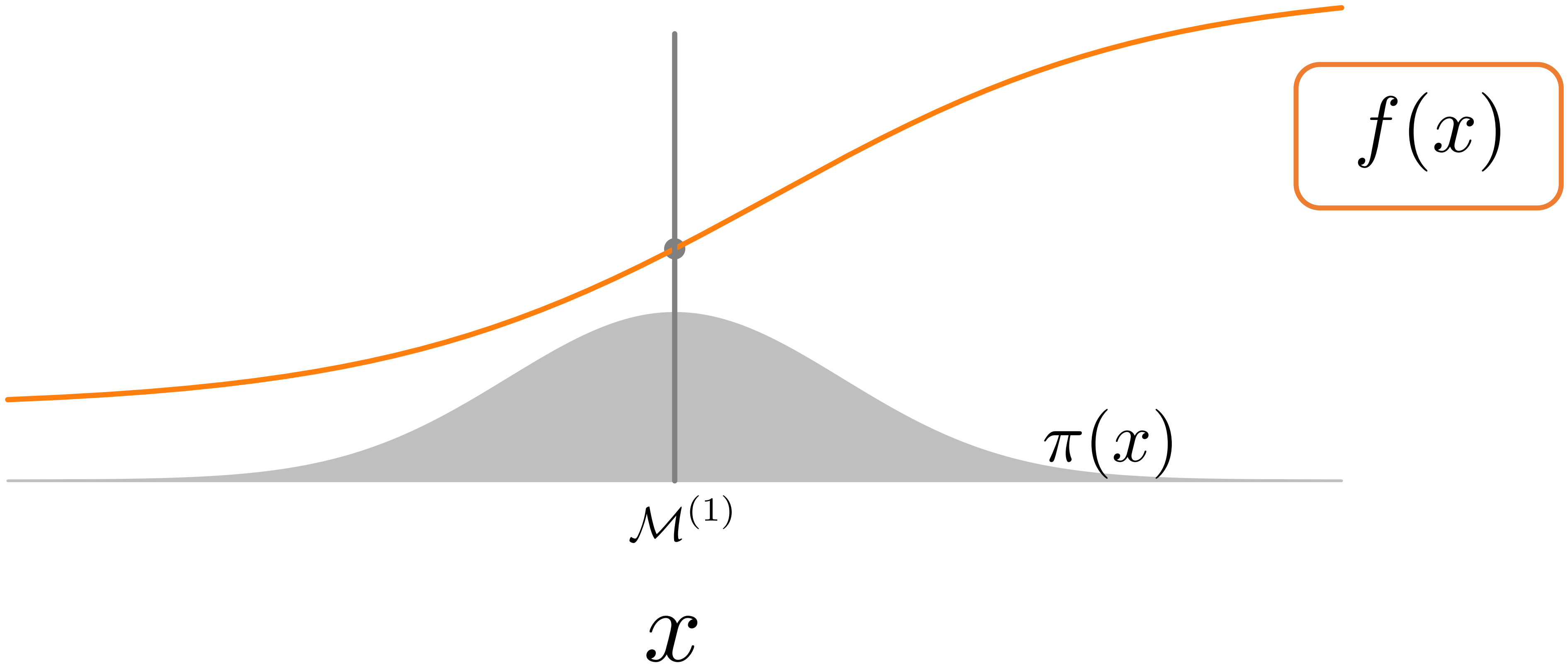
$$\begin{aligned} \mathbb{E}[f(X)] &= \mathbb{E}\left[f(x_0) + (X - x_0)^\intercal \frac{\partial f}{\partial x}(x_0) + R_{x_0}(X)\right] \\ &= \underbrace{f(x_0) + \left[\mathcal{M}^{(1)}_{x_0}\right]^\intercal \frac{\partial f}{\partial x}(x_0)}_{\text{constant}} + \underbrace{\mathbb{E}\left[R_{x_0}(X)\right]}_{\text{Monte Carlo}} \end{aligned}$$

shifted the variance to the residual term

Gaussian Example

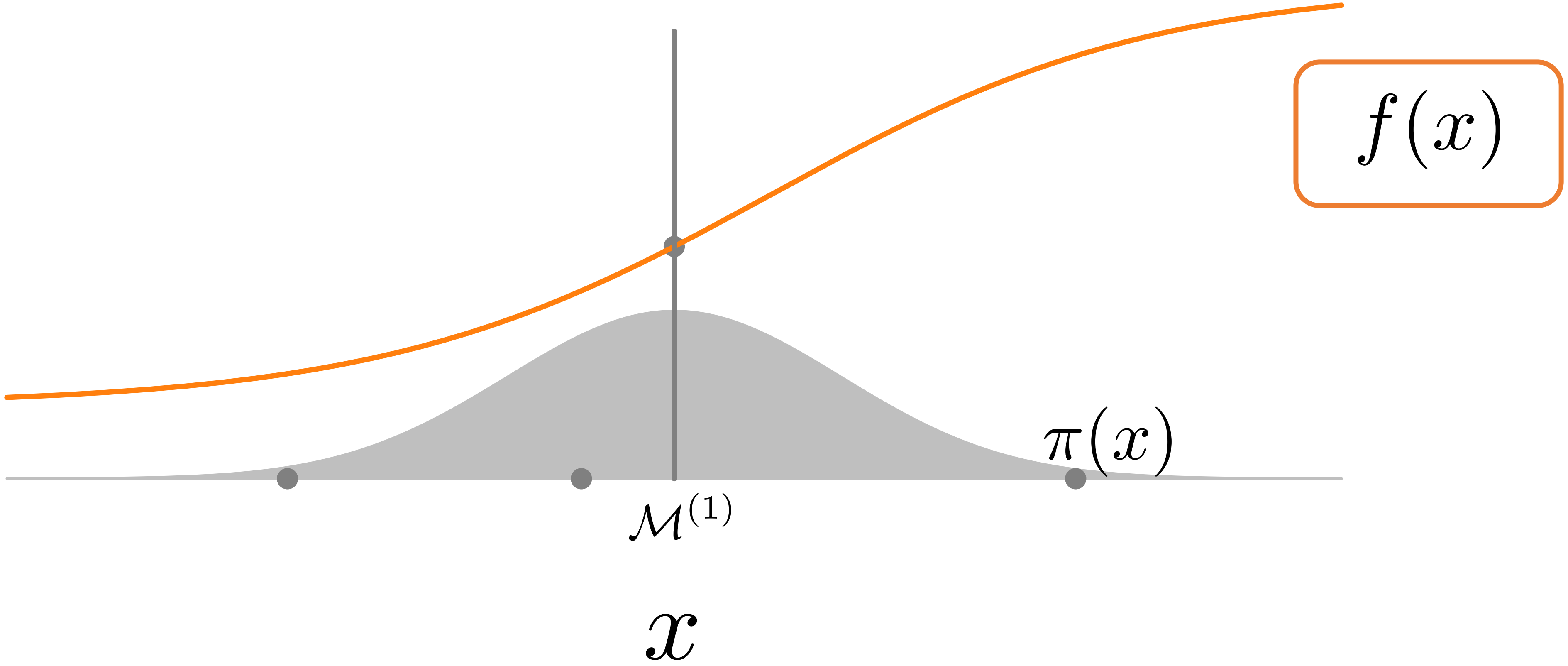


Gaussian Example



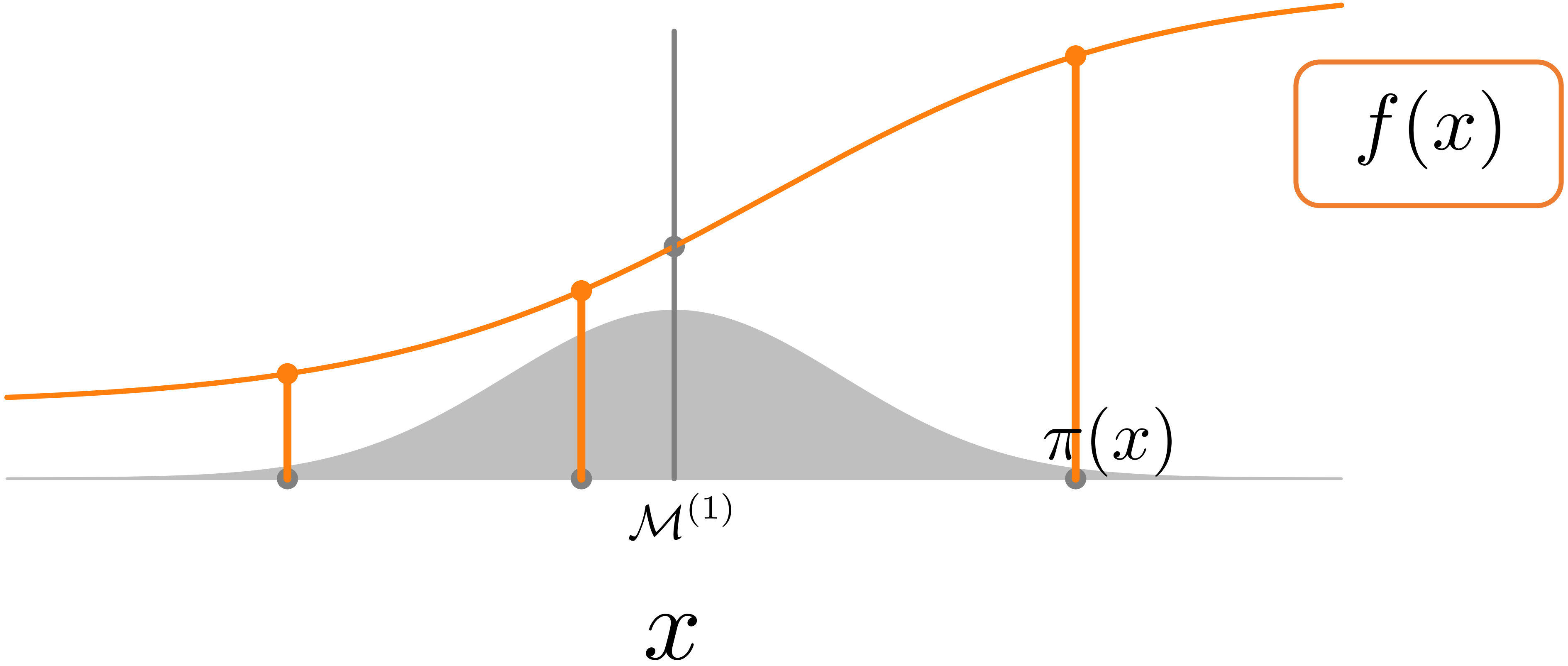
$$\mu_f = \mathbb{E}[f(X)]$$

Gaussian Example



$$\mu_f = \mathbb{E}[f(X)]$$

Gaussian Example

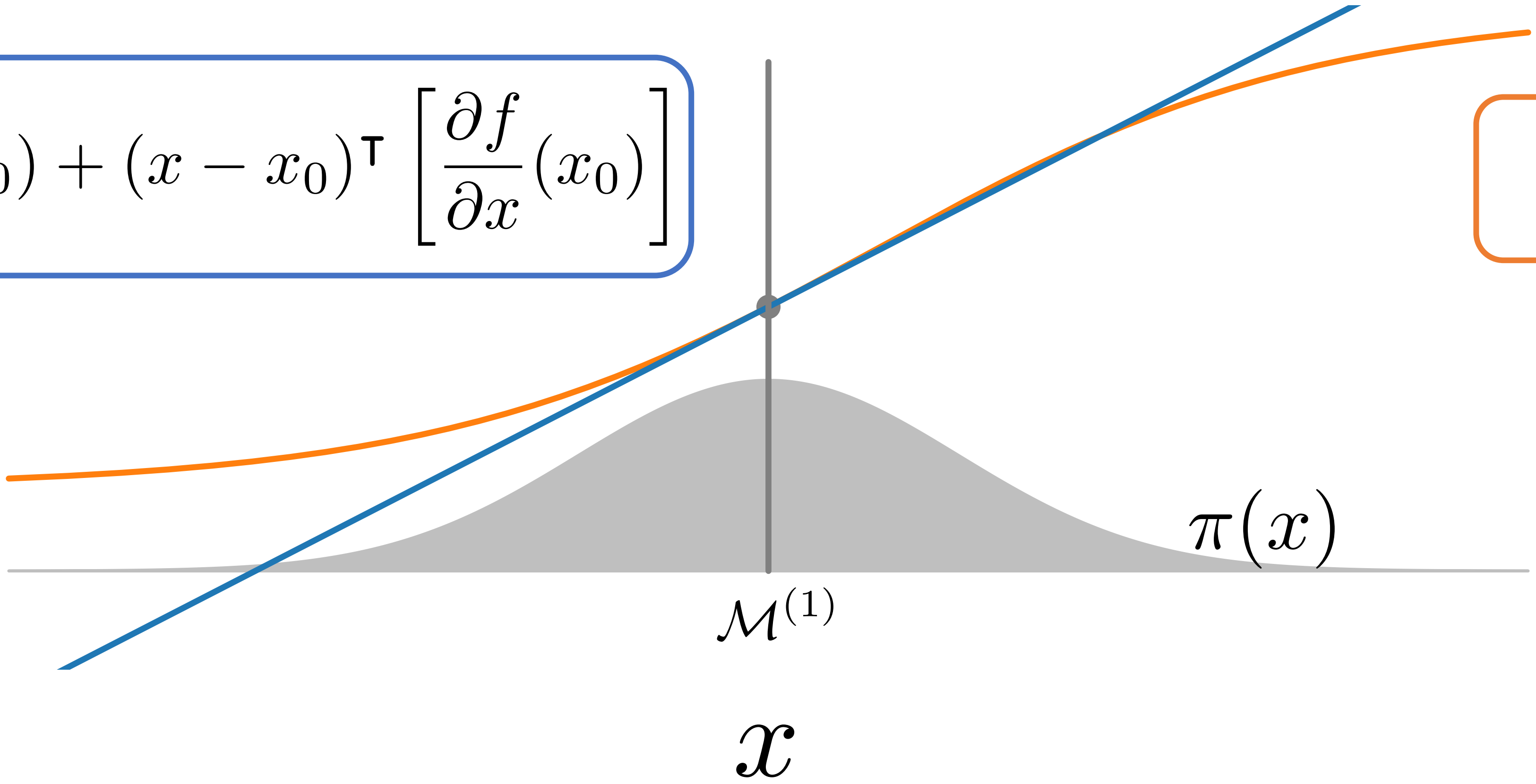


$$\mu_f = \mathbb{E}[f(X)]$$

Gaussian Example

$$f_{x_0}^{(1)}(x) = f(x_0) + (x - x_0)^\intercal \left[\frac{\partial f}{\partial x}(x_0) \right]$$

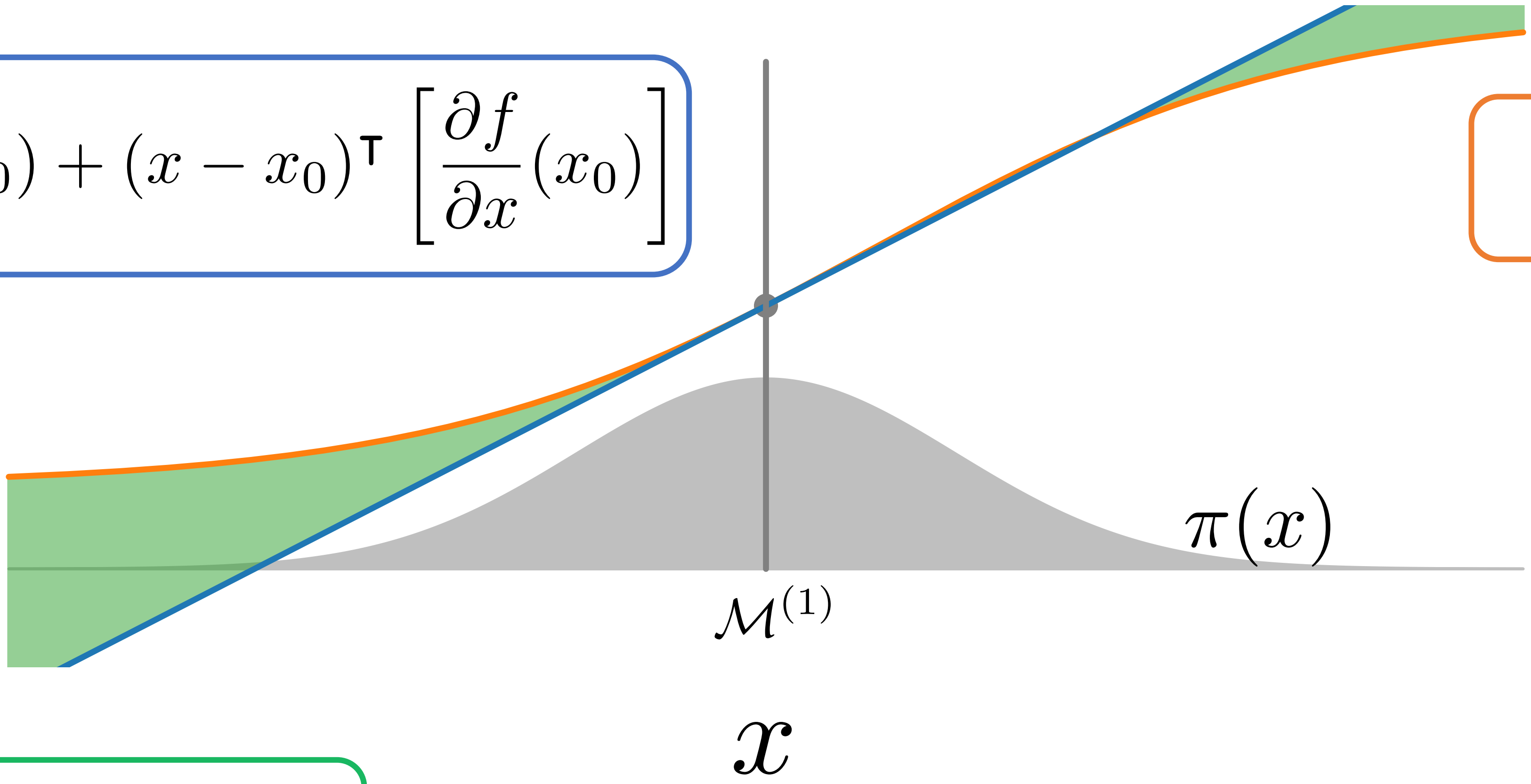
$$f(x)$$



$$\mu_f = \mathbb{E}[f(X)]$$

Gaussian Example

$$f^{(1)}_{x_0}(x) = f(x_0) + (x - x_0)^\intercal \left[\frac{\partial f}{\partial x}(x_0)\right]$$

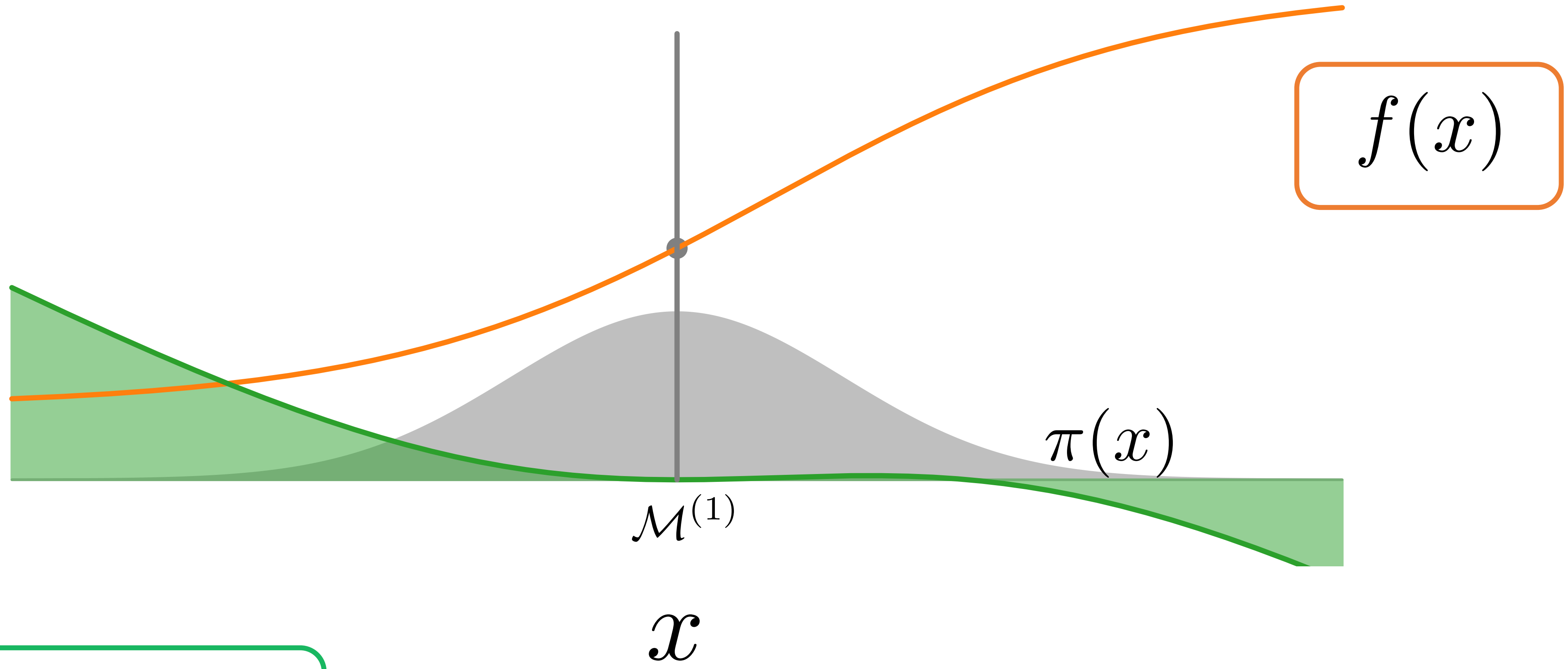


$$f(x)$$

$$R^{(1)}_{x_0}(x) = f(x) - f^{(1)}_{x_0}(x)$$

$$\mu_f = \mathbb{E}[f(X)]$$

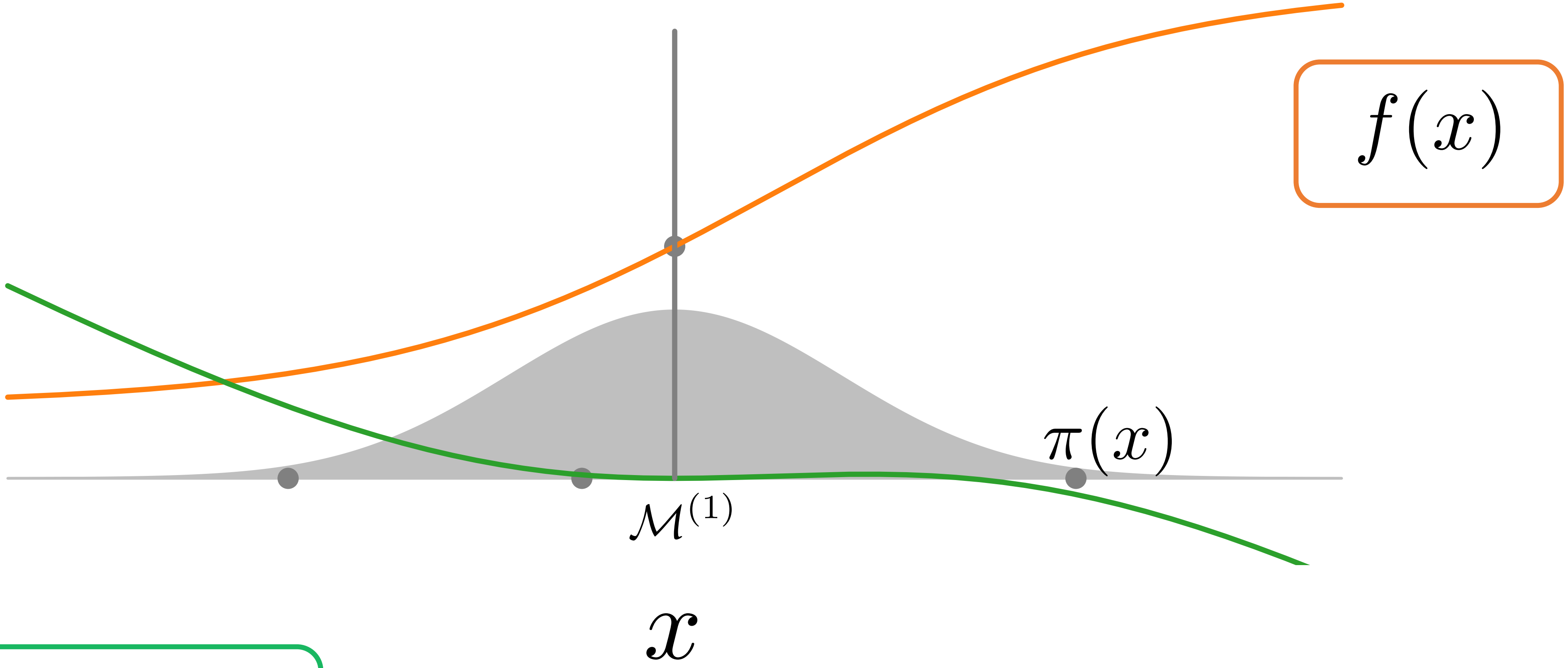
Gaussian Example



$$R^{(1)}_{x_0}(x) = f(x) - f^{(1)}_{x_0}(x)$$

$$\mu_f = \mathbb{E}[f(X)]$$

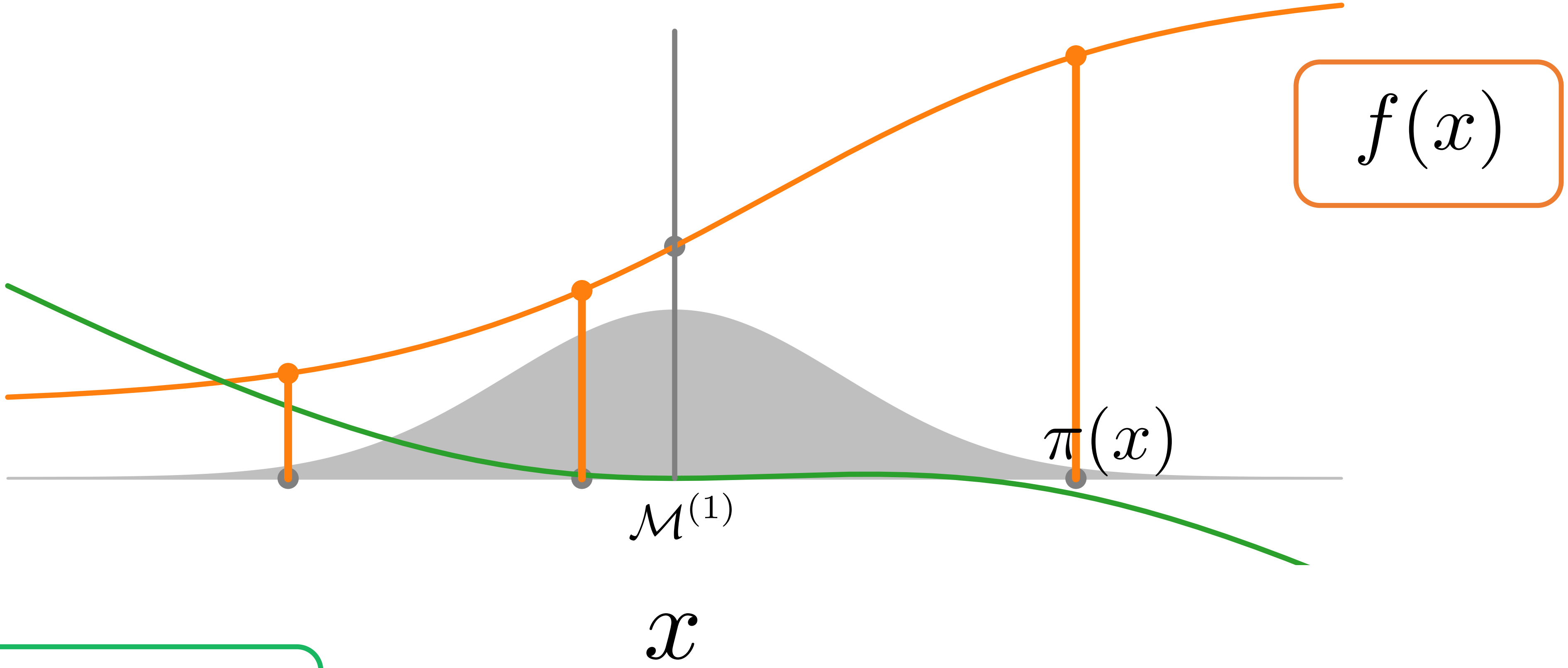
Gaussian Example



$$R_{x_0}^{(1)}(x) = f(x) - f_{x_0}^{(1)}(x)$$

$$\mu_f = \mathbb{E}[f(X)]$$

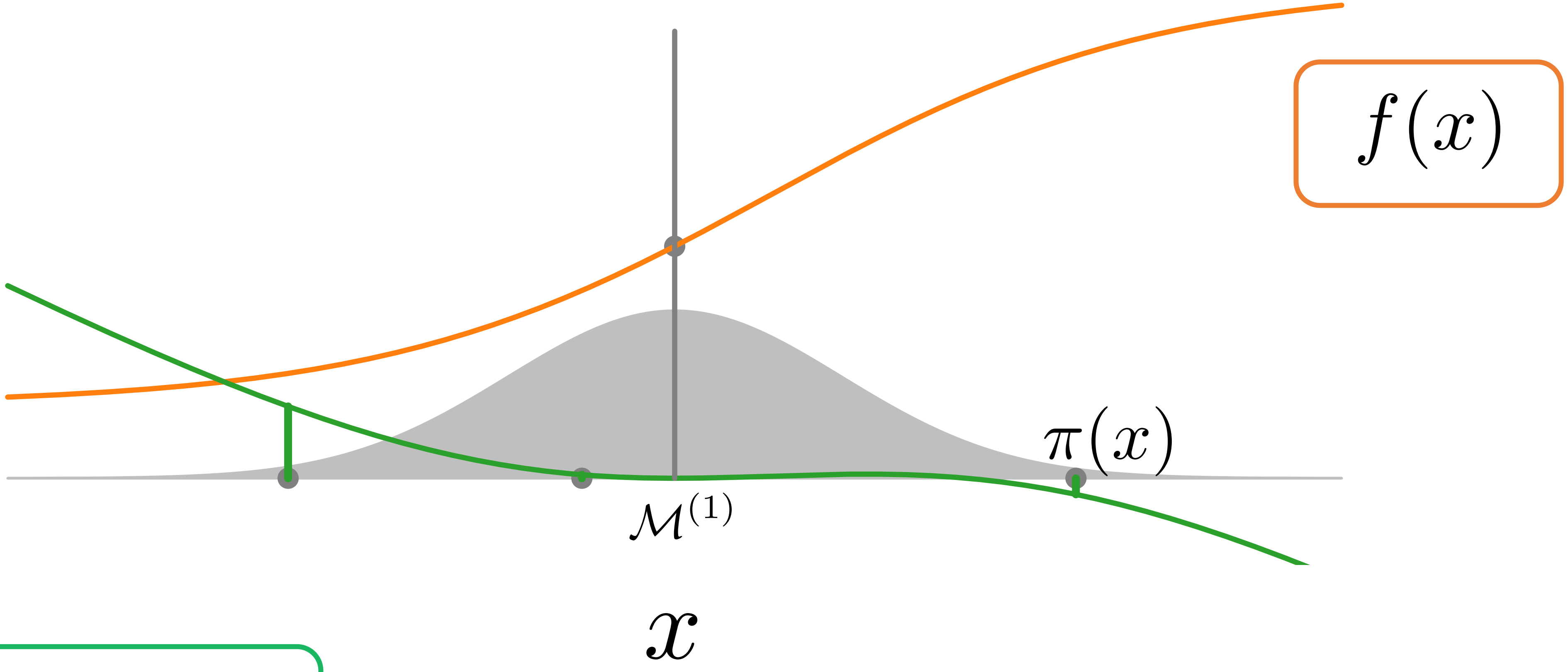
Gaussian Example



$$R_{x_0}^{(1)}(x) = f(x) - f_{x_0}^{(1)}(x)$$

$$\mu_f = \mathbb{E}[f(X)]$$

Gaussian Example



$$R^{(1)}_{x_0}(x) = f(x) - f^{(1)}_{x_0}(x)$$

$$\mu_f = \mathbb{E}[f(X)]$$

Taylor Residual Estimators

Assume we can compute …

$$\boxed{f(x)}$$

$$\boxed{R^{(1)}_{x_0}(x) = f(x) - f^{(1)}_{x_0}(x)}$$

$$\boxed{f^{(1)}_{x_0}(x) = f(x_0) + (x - x_0)^\intercal \left[\frac{\partial f}{\partial x}(x_0)\right]}$$

$$x \sim \pi$$

TRE 1

$$\hat{\mu}^{(1)}_f = \underbrace{f(x_0) + \left[\mathcal{M}^{(1)}_{x_0}\right]^\intercal \frac{\partial f}{\partial x}(x_0)}_{\text{constant}} + \underbrace{R^{(1)}_{x_0}(x)}_{\text{random}}$$

Taylor Residual Estimators

Assume we can compute …

$$\boxed{f(x)} \quad \boxed{R_{x_0}^{(M)}(x) = f(x) - f_{x_0}^{(M)}(x)}$$

$$\boxed{f_{x_0}^{(M)}(x) = f(x_0) + \sum_{m=1}^{M} \frac{1}{m!}(x - x_0)^m \frac{\partial^m f}{\partial x^m}(x_0)}$$

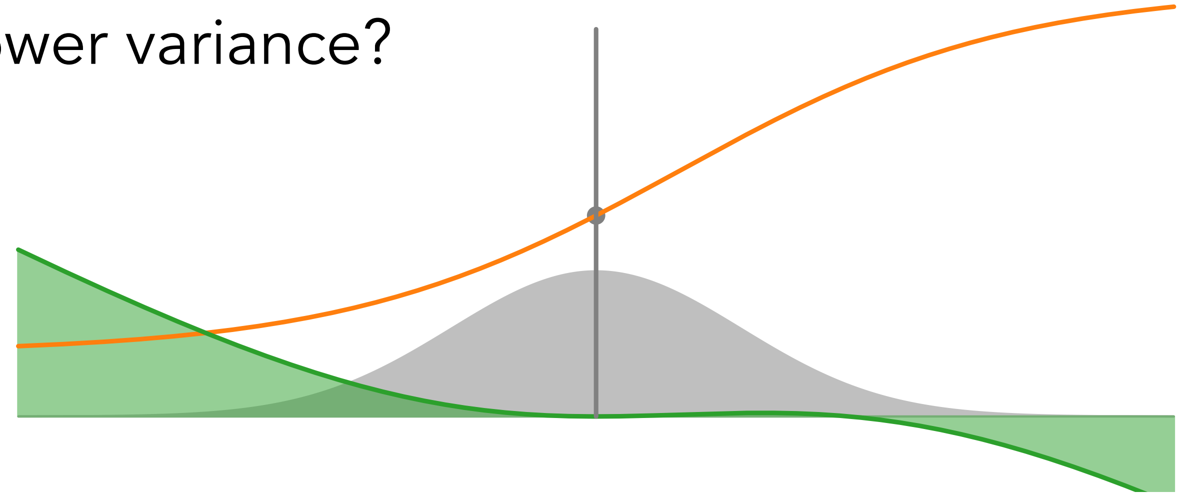
$$x \sim \pi$$

TRE M

$$\hat{\mu}_f^{(M)} = \underbrace{f(x_0) + \sum_m \frac{1}{m!} \mathcal{M}_{x_0}^{(m)} \frac{\partial^m f}{\partial x^m}(x_0)}_{\text{constant}} + \underbrace{R_{x_0}^{(M)}(x)}_{\text{random}}$$

Taylor Residual Estimators: Variance

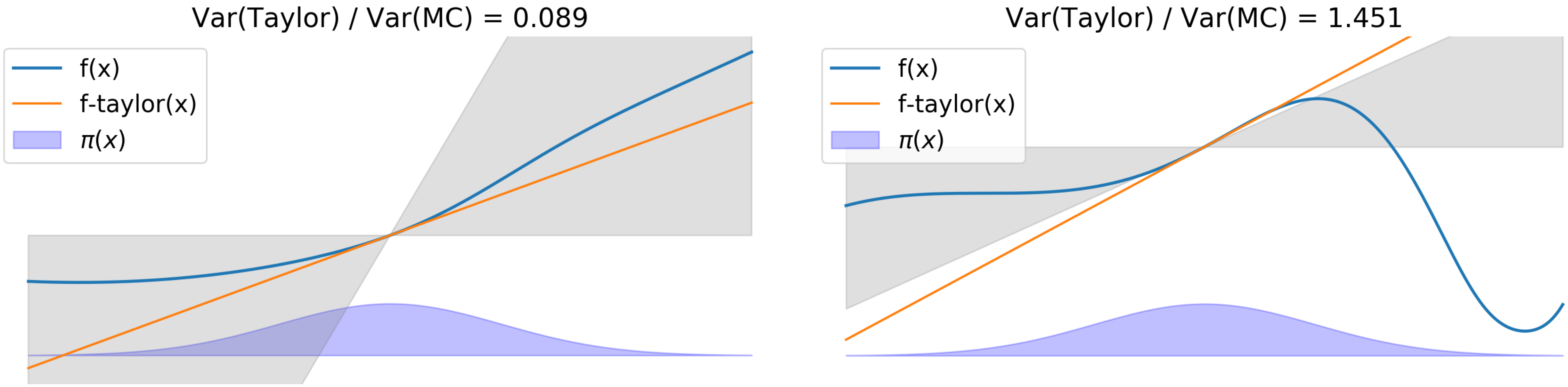
When will TREs have lower variance?



Sufficient condition (first order) $$\underbrace{\left|\frac{\partial f}{\partial x}(x_0)\right|}_{\text{Local linear approx}} \leq 2 \underbrace{\left|\boxed{\mathbb{V}(X)}^{-1}\boxed{\mathbb{C}(X, f(X))}\right|}_{\text{Population least squares coefficient}}$$

Taylor Residual Estimators: Variance

When will TREs have lower variance?



Variance reduction

Variance increase

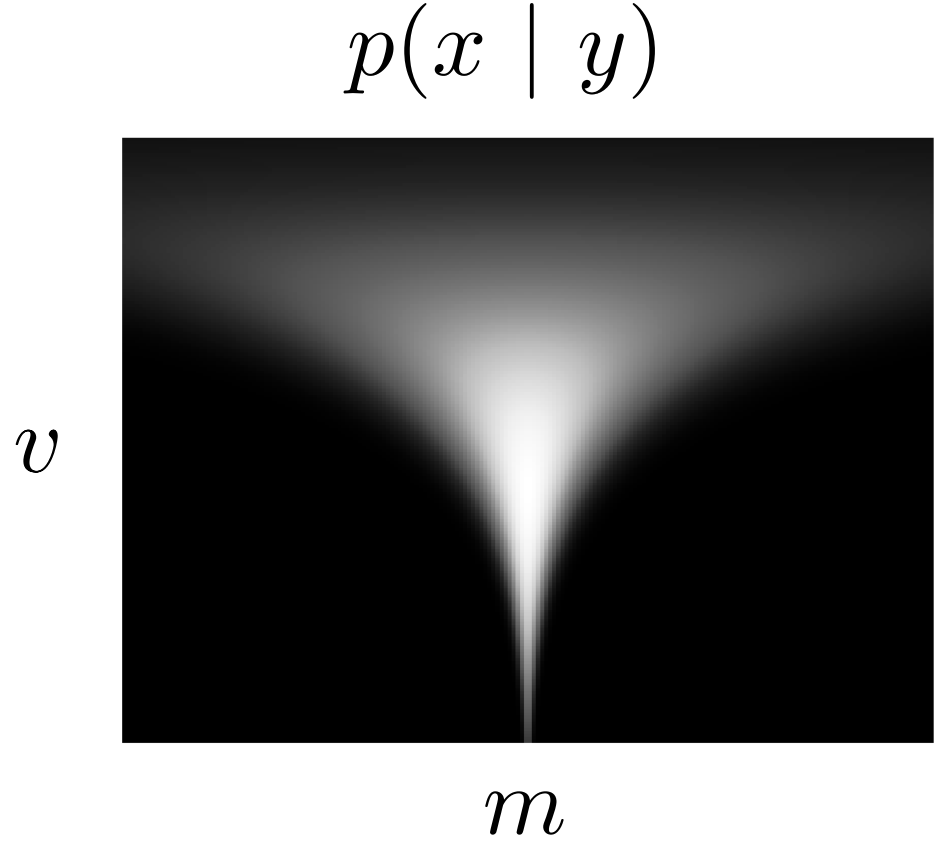
Experiments

$$x = [m, v]$$

$$v \sim \mathcal{N}(0, 3^2)$$

$$y \sim \mathcal{N}(m, \exp(v/2))$$

$$\dim(x) = 20$$



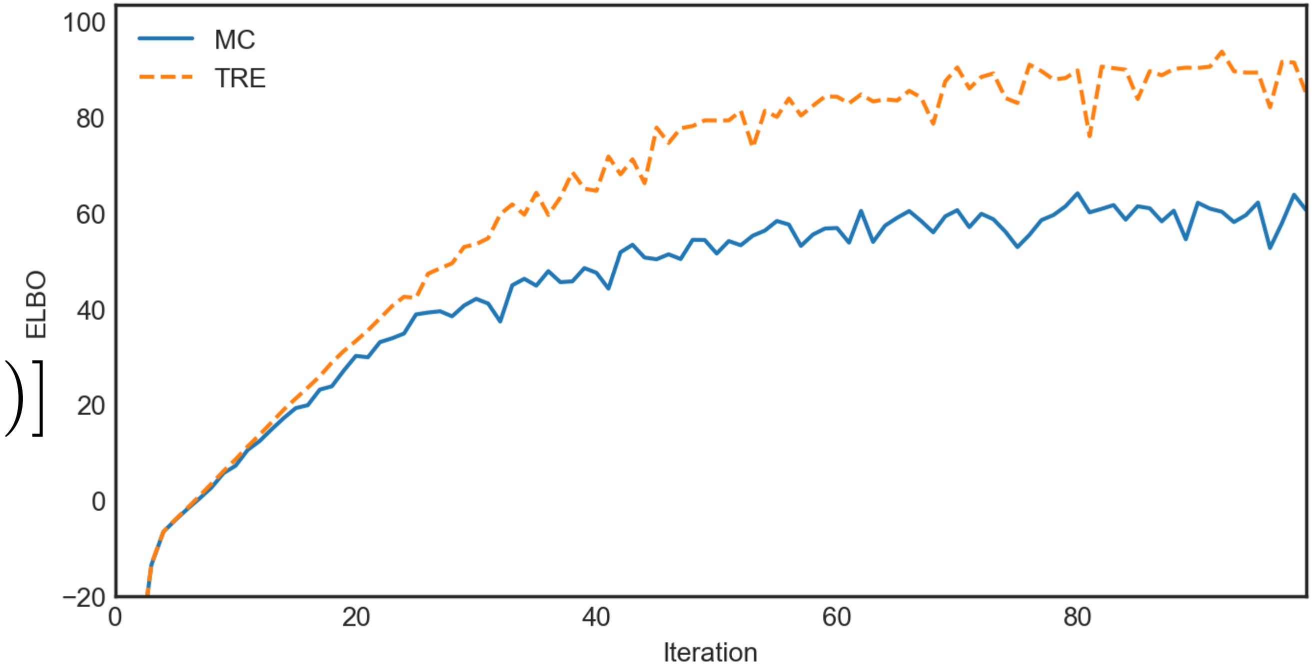
Experiments

Gaussian ELBO

$$q(x;\boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_\sigma)$$

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{X\sim q}\left[\ln \pi(X,\mathcal{D}) - \ln q(X;\boldsymbol{\lambda})\right]$$

TRE estimator of the ELBO: 320x lower variance than Monte Carlo at initialization; comparable at convergence



Optimization comparison

Experiments

Normalizing Flows ELBO (Planar Flow)

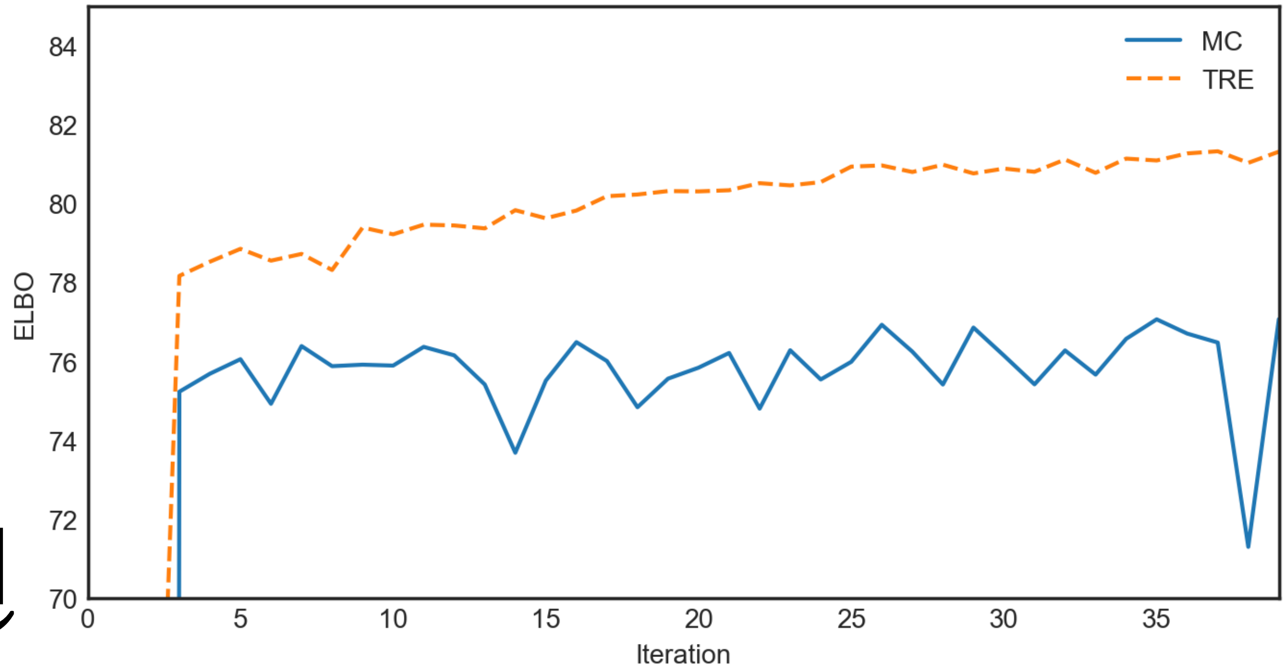
$$x_0 \sim \mathcal{N}(0, I_D)$$

$$x_1 = \phi(x_0; \boldsymbol{\lambda}_1)$$

$$\dots$$

$$x = x_L = \phi(x_{L-1}; \boldsymbol{\lambda}_L)$$

$$\mathcal{L}(\boldsymbol{\lambda}) = \underbrace{\mathbb{E}_q[\ln \pi(X, \mathcal{D})]}_{\text{model term}} - \underbrace{\mathbb{E}[\ln q(X; \boldsymbol{\lambda})]}_{\text{entropy term}}$$

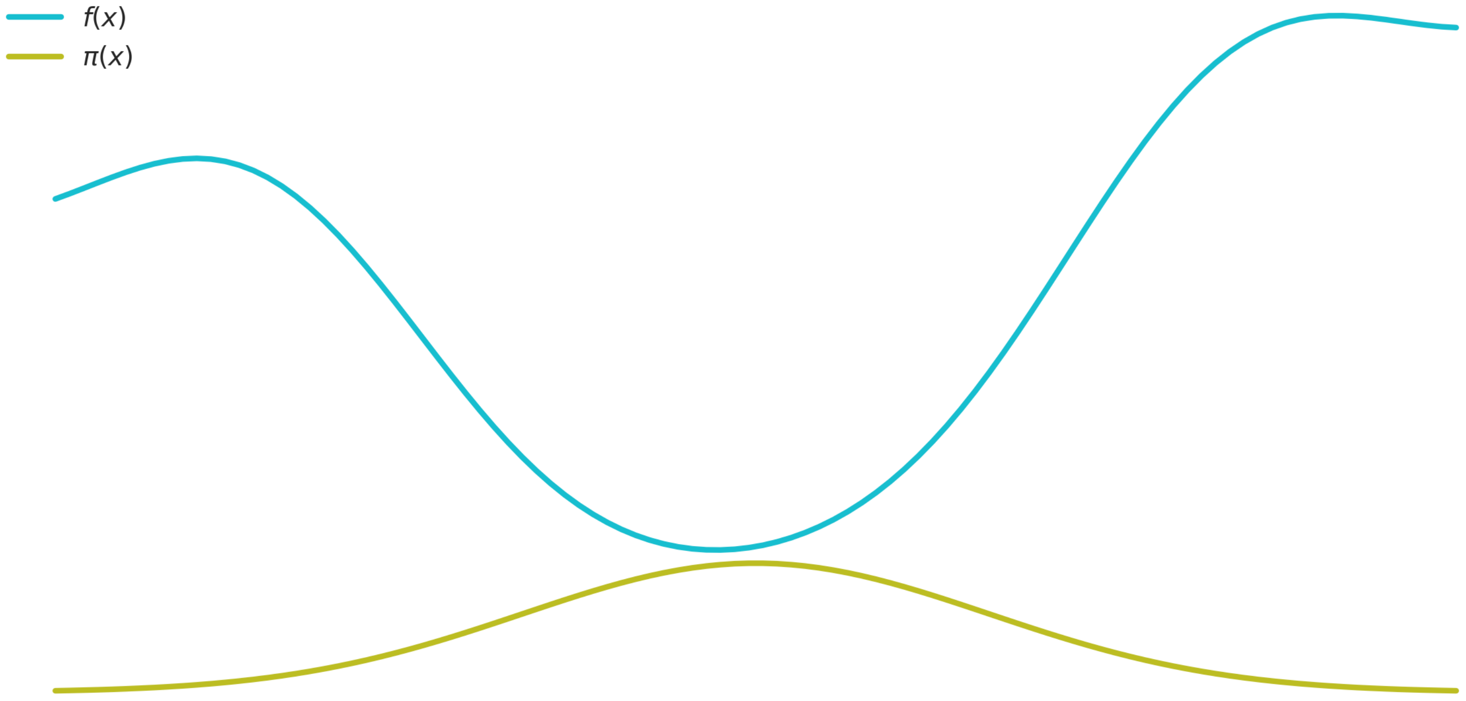


Optimization comparison

TRE estimator of the ELBO: 40x lower variance than Monte Carlo at initialization; 2x lower at convergence

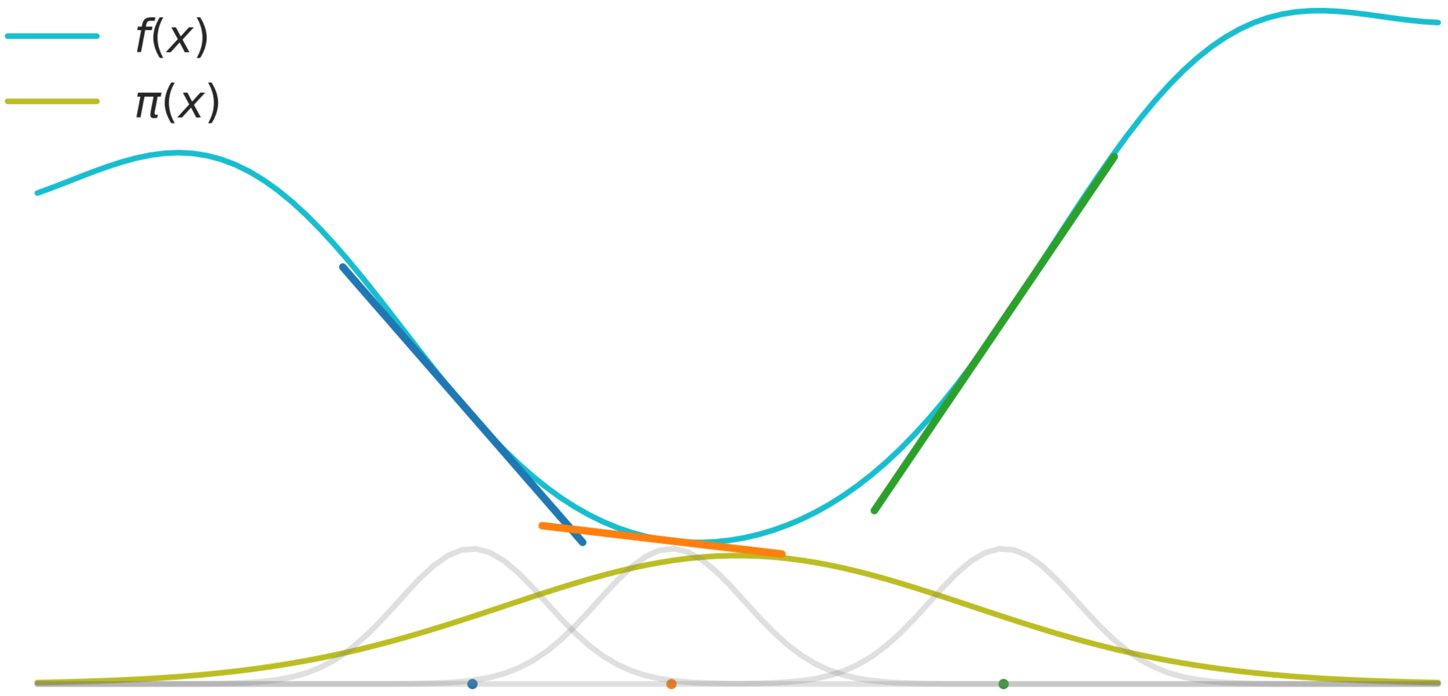
Future/Ongoing Work

What if the local Taylor approximation fails?



Future/Ongoing Work

What if the local Taylor approximation fails?



Thanks!

Questions? Comments?

acm@seas.harvard.edu
http://andymiller.github.io
twitter: @_amiller_